Dean of Graduate Studies
Al-Quds University

Comparison of Data Mining and Statistical Techniques for Prediction Model

Amjad A. M. Harb

M.Sc. Thesis

Jerusalem-Palestine

1433 / 2012

Comparison of Data Mining and Statistical Techniques for Prediction Model

Prepared By:
Amjad A. M. Harb

B.Sc.: Computer Science-Al-Quds University-Palestine

Supervisor: Dr. Rashid Jayousi

A thesis Submitted in Partial fulfillment of requirements for the degree of Master of Computer Science / Department of Computer Science / Faculty of Graduate Studies – Al-Quds University.

1433 / 2012

Al-Quds University
Deanship of Graduate Studies
Computer Science Department

Thesis Approval

Comparison of Data Mining and Statistical Techniques for Prediction Model

Prepared By: Amjad A. M. Harb
Registration No: 20911925

Supervisor: Dr. Rashid Jayousi

Master thesis submitted and accepted. Date: / /2012
The names and signatures of the examining committee members are as follows:

1- Head of Committee: Dr. Rashid Jayousi	Signature: ……..…..
2- Internal Examiner: Dr. Nidal Kafri	Signature: ……..……..
3- External Examiner: Dr. Yousef Abuzir	Signature: ……..……..

Jerusalem – Palestine

1433 / 2012

Dedication

This work is dedicated…

To my parents, for their love, endless support and encouragement…

To my beloved wife, without her caring support it would not have been possible…

To my sons and daughters …. Majd, Abdel-Munem, Omar, Mohammad, and Sara…

To my brothers, sisters, friends and colleagues…

To all of you I say a big

"Thank you" for being example of love and care.

Amjad A. M. Harb

Declaration

I certify that this thesis submitted for the degree of Master, is the result of my own research, except where otherwise acknowledged, and that this study (or any part of the same) has not been submitted for a higher degree to any other university or institution.

Signed……..

Amjad A. M. Harb

Date: / /2012

Acknowledgments

First and foremost Praise be to the Almighty Allah, Lord of all creatures, the Most Gracious, Most Merciful, for His graces and blessings throughout all my life. Without Him, everything is nothing.

My sincere thanks for my supervisor Dr. Rashid Jayousi, for his sincere efforts, interest and time he have kindly spent to guide my research.

I am extremely grateful and indebted to my colleagues in PCBS, Qais Hasiba and Nayef Abed, for helping me during the study.

I am very grateful to all professors at Al-Quds University-Computer Science department, for the time they spent to teach me.

As my study was partially funded by my employer, the Palestinian Central Bureau of Statistics, special thanks to PCBS represented by its president Mrs. Ola Awad and all the employees in PCBS.

Finally, and most importantly, I would like to thank my wife. Her support, encouragement, quiet patience and unwavering love were undeniable.

Abstract

The aim of this research is to perform a comparison study between statistical and data mining modeling techniques. These techniques are statistical Logistic Regression, data mining Decision Tree and data mining Neural Network. The performance of these prediction techniques were measured and compared in terms of measuring the overall prediction accuracy percentage agreement for each technique and the models were trained using eight different training datasets samples drawn using two different sampling techniques. The effect of the dependent variable values distribution in the training dataset on the overall prediction percent and on the prediction accuracy of individual "0" and "1" values of the dependent variable values was also experimented. For a given data set, the results shows that the performance of the three techniques were comparable in general with small outperformance for the Neural Network. An affecting factor that makes the percent prediction accuracy varied is the dependent variable values distribution in the training dataset, distribution of "0" and "1". The results showed that, for all the three techniques, the overall prediction accuracy percentage agreement was high when the dependent variable values distribution ratio in the training data was greater than 1:1 but at the same time they, the techniques, fails to predict the individual dependent variable values successfully or in acceptable prediction percent. If the individual dependent variable values needed to be predicted comparably, then the dependent variable values distribution ratio in the training data should be exactly 1:1.

دراسة مقارنة الأداء والكفاءة بين تقنيات التنقيب عن البيانات والأساليب الإحصائية في تصميم نماذج التنبؤ

إعداد: امجد عبد المنعم محمود حرب

إشراف: د. رشيد الجيوسي

ملخص:

هدف هذه الدراسة هو إجراء مقارنة الكفاءة والفعالية بين الوسائل الإحصائية وتقنيات التنقيب عن البيانات لبناء نماذج التصنيف والتنبؤ العلمي. الخوارزميات والوسائل والتقنيات التي تمت دراستها ومقارنة أدائها هي الانحدار اللوجستي الإحصائي، وتقنيتي التنقيب عن البيانات شجرة القرار والشبكة العصبية. تم قياس أداء هذه التقنيات ومقارنتها بالاعتماد على مقياس مشترك وهو النسبة المئوية الشاملة لدقة التنبؤ لكل تقنية. تم تدريب نماذج هذه التقنيات باستخدام ثمانية عينات من بيانات التدريب تم سحبها بالاعتماد على تقنيتي سحب عينات إحصائية. تم أيضا فحص تأثير توزيع قيم المتغير التابع في بيانات تدريب خوارزميات التنبؤ المذكورة وذلك على مستوى النسبة المئوية الشاملة لدقة التنبؤ لكل تقنية وأيضا على مستوى النسبة المئوية لدقة التنبؤ لقيم المتغير التابع الفردية "0" و "1" لكل تقنية. أظهرت النتائج أن أداء التقنيات الثلاثة كانت بشكل عام متقاربة وقابلة للمقارنة مع تفوق بسيط لخوارزمية الشبكات العصبية. تم تحديد عنصر مؤثر على اختلاف وتفاوت دقة النسبة المئوية للتنبؤ وهذا العنصر هو توزيع قيم المتغير التابع في بيانات تدريب النماذج، أي توزيع "0" و "1". كما أظهرت النتائج أيضا أن النسبة المئوية لدقة التنبؤ الشامل للتقنيات الثلاثة

كانت مرتفعة عندما كانت نسبة توزيع قيم المتغير التابع في بيانات التدريب أكبر من 1:1 ولكن في الوقت نفسه فشلت الخوارزميات والتقنيات قيد الدراسة في التنبؤ بالقيم الفردية للمتغير التابع بنجاح أو بنسبة تنبؤ مقبولة. في التطبيقات باستخدام هذه التقنيات إذا كان الهدف هو الحصول على تنبؤ بنسبة مئوية عالية لقيم المتغير التابع الفردية وأن تكون النسبة المئوية للتنبؤ بالقيمتين متقاربة فانه يجب أن تكون نسبة توزيع قيم المتغير التابع في بيانات التدريب بالضبط تساوي 1:1.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter One

Introduction

An important and challenging area of research nowadays is machine learning. Historical data was analyzed using several ways for hidden knowledge extraction that can help in decision making, and this is called *Knowledge Discovery* or *Data Mining*. The popular goal from data mining is prediction and the popular data mining technique used for prediction is classification. Classification can be accomplished statistically or by data mining methods. [1]

Comparison studies in prediction techniques performance are interesting topics for many researchers. For example a comparative study by Lahiri R. [1] compared the performance of three statistical and data mining techniques on Motor Vehicle Traffic Crash dataset, resulted that the data information content and dependent variable distribution is the most affecting factor in prediction performance. Another study by Delen D. et al. [2] targeted data mining methods comparison as a second objective in the study, while the main objective was to build the most accurate prediction model in a critical field, breast cancer survivability. In the same domain, Artificial Intelligence in Medicine, Bellaachia A. et al. [3] continued the work done by Delen D. et al. [2] and improved the research tools especially the dataset. An important application area that exploited data mining techniques heavily was the network security. Panda M. et al. [4] also performed a comparative study

to identify the best data mining technique in predicting network attacks and intrusion detection. Also the data contents and characteristics revealed as an affecting factor on the data mining and prediction algorithms performance.

The work in this research depended on the methodology of Lahiri R. [1] and extended the experiment further to investigate the effect of the dependent variable values distribution in the training data on the prediction accuracy of the prediction techniques, viz., Logistic Regression, Neural Network and Decision Tree in addition to the main objectives of the research to compare the overall prediction accuracy percent performance of the prediction techniques over different training datasets samples drawn using two different sampling methods.

1.1 Motivation

As data mining is a new area of research and we work in the same field, producing and disseminating official statistics, we have a large interest in this field especially prediction and data visualization. So identifying the active and suitable prediction techniques is essential and highly useful in our work.

1.2 Objectives

In this research we will continue on the work of Lahiri R. [1] to perform a comparison on the same statistical and data mining techniques, viz., Logistic Regression, Neural Network and Decision Tree but with more accurate data content and quality, as Lahiri's future work recommendation, which can be achieved by selecting more precise predictors that significantly define and affect the output. In other words, we intended to ask for the help of the statistician domain experts to select the independent variables and then apply a correlation test to select the most correlated variables, as predictors, to the dependent variable and examine the prediction accuracy rates using the aforementioned prediction

techniques. The effect of training data sampling method and sample size will be explored and highlighted. The overall prediction percentage agreement will be the main performance metric. A secondary objective of the research is to measure and identify the effect of the dependent variable values distribution, "0" and "1", in the dataset on the overall prediction accuracy and on the prediction accuracy of "0" and "1" individually, using the three prediction techniques.

The experiment will exploit a historical dataset about the "Palestinian Expenditure and Consumption Survey" produced by the Palestinian Central Bureau of Statistics (PCBS) [26]. The dependent variable will be the household's "Level of Poverty", that have the values: "0" as "not poor" and "1" as "poor".

It is worth mentioning that as a result of this research, two scientific research papers were published in the proceedings of the International Conference on Information and Communications Technology (ICICT'2012) [23] and the 13th International Arab Conference on Information Technology (ACIT'2012) [24]. A third scientific paper was submitted to the 6th International Conference on Information Technology (ICIT'2013) [25].

This thesis is organized as follows: The literature and related work will be discussed in chapter two. A background of the research including a description about the data and the techniques and methods used in the research was presented in chapter three. The research methodology followed to perform the experiment was presented in chapter four. Experimental results are presented and discussed in chapter five. Finally, the conclusion was given in the last chapter six.

Chapter Two

Literature Review

Many studies have been done across countries on data mining. Applications of data mining were used in a large number of fields, especially for business and medical purposes.

Data mining is a new technology field and it is important and very helpful in predicting and detecting underlying patterns from large volumes of data. Many researches were published, comparing results of data mining algorithms in several areas. A research by Rochana Lahiri (2006) performed a performance comparison of several data mining and statistical techniques for classification model. She used a database from Louisiana Motor Vehicle Traffic Crash. The performance was measured in terms of the classification agreement percent. The effect of Decision Tree, Neural Network, and Logistic Regression models for different sample sizes and sampling methods on three sets of data had been investigated. The study concluded that a very large training dataset is not required to train a Decision Tree model or a Neural Network model or even for Logistic Regression model to obtain high classification accuracy and the overall performance reached a steady value at the sample size of 1000, irrespective of the total population size. The information content of a training dataset is an important factor influencing classification accuracy and is not governed by the size of the dataset. Another important result was that the sampling method

has not affected the classification accuracy of the models. She also stated that the overall classification accuracy of the all three methods were very much comparable and no one method over performed any other. She tried to find the effect of the "0"s and "1"s distribution of dependent variable values in the dataset but because the data was very skewed, she failed to do this. As a future work, the study recommends to apply the same study on a dataset were the relationships between the dependent variable and the independent variables are more rigid. i.e.: to select predictors that strongly describe the dependent attribute, and to study the effect of the distribution "0"s and "1"s that represent dependent variable values. [1]

The data mining methods comparison were targeted as a second objective in some studies that mainly aimed to develop a prediction model in a critical fields, like medicine, by investigating several data mining methods, intending to get the model that have the highest prediction accuracy. This type of studies has been addressed by Delen D. et al. (2005) in the context of predicting breast cancer survivability. Multiple prediction models, using Artificial Neural Networks, Decision Trees, and Logistic Regression, for breast cancer survivability using a large dataset had been developed. The comparison among the three models had been conducted depending on measuring three prediction performance metrics: classification accuracy, sensitivity and specificity. The k-Fold cross-validation test was used to minimize the bias associated with the random sampling of the training and missing data. The results of the study showed that the Decision Tree (C5) preformed the best of the three models evaluated. Sensitivity analysis, which provides information about the relative importance of the input variables in predicting the output field, was applied on Neural Network models and provided them with the prioritized importance of the prediction factors used in the study. [2]

Another related study in medicine by Bellaachia A. et al. (2006) also in the context of predicting breast cancer survivability. The researchers took the study of Delen D. et al. [2] as the starting point with the same dataset source but with a newer version and different set of data mining techniques. For modeling and comparison, three data mining techniques had been investigated: the Naïve Bayes, the back-propagated Neural Network, and the C4.5 Decision Tree algorithms. The main goal was to have a prediction model with high prediction accuracy, besides high precision and recall metrics for patients' data retrieval. They used other performance metrics: specificity and sensitivity to compare the prediction models. The results presented that C4.5 algorithm has a much better performance than the other two techniques. The obtained results differed from the study of Delen D. et al. [2] due to the facts that they used a newer database (2000 vs. 2002), a different pre-classification (109,659 and 93,273 vs. 35,148 and 116,738) and different toolkits (industrial grade tools vs. Weka). [3]

In network security, data mining techniques used heavily in predicting network intrusion detection systems to protect computing resources against unauthorized access. Several studies were performed in this area and some of them addressed the prediction performance comparison of different data mining techniques like the study by Panda M. et al. (2008). A dataset of 10% KDDCup'99 intrusion detection has been generated and used in the experiment. Three popular data mining algorithms had been used in the experiment: Decision Trees ID3, J48 and Naïve Bayes. The prediction performance metrics used in the study were the time taken to build the model and the prediction error rate. For the evaluation of prediction error rate, the 10-fold cross validation test was used. As a result of the experiment, the Decision Trees had proven their efficiency in both generalization and detection of new attacks more than the Naïve Bayes. But this maybe dependence on the

contents and characteristics of the data which allows single algorithm to outperform others. [4]

Amooee G. et al. (2011) used data mining techniques to identify defective parts manufactured in an industrial factory and to maintain high quality products. A data of 1000 records was collected from the factory and 10% (100 records) of the data was about a defective parts. Prediction accuracy and processing time of the prediction techniques were the comparison performance metrics. The results showed that SVM and Logistic regression prediction algorithms has the best processing time with high overall prediction accuracy. The decision tree with its tree different branching algorithms (CRT, CHAID, and QUEST) achieved the highest prediction accuracy rates but needed more time. Neural Network achieved the least prediction accuracy rate with medium processing time. [5]

Data mining concept was the most appropriate to the study of student retention from sophomore to junior year than the classical statistical methods. This was one main objective of the study addressed by Ho Yu C. et al. (2010) in addition to another objective that identifying the most affecting predictors in a dataset. The statistical and data mining methods used were classification tree, multivariate adaptive regression splines (MARS), and Neural Network. The results showed that transferred hours, residency, and ethnicity are crucial factors to retention, which differs from previous studies that found high school GPA to be the most crucial contributor to retention. In Ho Yu C. et al. research, the Neural Network outperformed the other two techniques. [6]

The prediction techniques RIPPER, decision tree, Neural Networks and support vector machine were used to predict cardiovascular disease patients. The performance comparison metrics were the Sensitivity, Specificity, Accuracy, Error Rate, True Positive Rate and

False Positive Rate. Kumari M. et al. (2011) study showed that support vector machine model outperforms the other models for predicting cardiovascular disease. [7]

The Neural Network was found to achieve better performance compared to the performance rates of Naive Bayes, K-NN, and decision tree prediction techniques in a study performed by Shailesh K R et. al. (2011) to predict the inpatient hospital length of stay in a super specialty hospital. [8]

The same result was seen that the Neural Network outperformed both the decision tree and linear regression models when the performance for the students' academic performance in the undergraduate degree program was measured by predicting the final cumulative grade point average (CGPA) of the students upon graduation. The correlation coefficient analysis was used to identify the relationship of the independent variables with the predictors. Ibrahim Z. et al. (2007) [9]

Social network data, using data mining techniques and the prediction error rates were the comparison metric, was studied by Nancy P. et al. (2011). The tree based algorithms such as RndTree, ID3, C-RT, CS-CRT, C4.5, CS-MC4 and the k-nearest neighbor (k-NN) algorithms were used in the study. The RndTree algorithm achieved least error rate and outperforms the other algorithms. [10]

C. Deepa et al. (2011) compared the prediction accuracy and error rates for the compressive strength of high performance concrete using MLP Neural Network, Rnd tree models and CRT regression. The results showed that Neural Network and Rnd tree achieved the higher prediction accuracy rates and Rnd tree outperforms Neural Network regarding prediction error rates. [11]

The Rand tree algorithm also outperforms the other algorithms, C4.5, C-RT, CS-MC4, decision list, ID3 and naïve bayes, in a study of vehicle collision patterns in road accidents

by S. Shanthi et al. (2011). Selection algorithms were used including CFS, FCBF, Feature Ranking, MIFS and MODTree, to improve the prediction accuracy. Feature Ranking algorithm was found the best in improving the prediction accuracy for all algorithms[12]. Table (2.1) presents a summary of the above literature review contributions.

Table 2.1: summary of literature review contributions according to area of research

Area of Research	Study Title	Author	Year	Main Contribution
Vehicle Collisions	Comparison of Data Mining and Statistical Techniques for Classification Model	Lahiri R.	2006	The study concluded that a sample training dataset of size 1000 records, irrespective of the total population size, can efficiently train a Decision Tree model or a Neural Network model or even for Logistic Regression model to obtain high classification accuracy. Also the sampling method has not affected the classification accuracy of the models.
	Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms	S.Shanthi et al.	2011	In vehicle collision patterns in road accidents the Rand tree algorithm outperformed the other algorithms, C4.5, C-RT, CS-MC4, decision list, ID3 and naïve bayes, Selection algorithms were used including CFS, FCBF, Feature Ranking, MIFS and MODTree, to improve the prediction accuracy. Feature Ranking algorithm was found the best in improving the prediction accuracy for all algorithms.
Social Network	A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data	Nancy P. et al.	2011	In Social network data, the tree based algorithms such as RndTree, ID3, C-RT, CS-CRT, C4.5, CS-MC4 and the k-nearest neighbor (k-NN) algorithm prediction performance were compared using the prediction error rates as comparison metric. The RndTree algorithm achieved least error rate and outperforms the other algorithms.
Network Security	A Comparative Study Of Data Mining Algorithms For Network Intrusion Detection	Panda M. et al.	2008	Decision Trees ID3, J48 and Naïve Bayes prediction performance was compared using the time taken to build the model and the prediction error rate as performance metrics. The Decision Trees had proven their efficiency in both generalization and detection of new network attacks more than the Naïve Bayes.

Area of Research	Study Title	Author	Year	Main Contribution
Medicine	Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods	Delen D. et al.	2005	The prediction accuracy performance comparison of Neural Networks, Decision Trees, and Logistic Regression showed that the Decision Tree (C5) preformed the best of the three models. Sensitivity analysis, which provides information about the relative importance of the input variables in predicting the output, was applied on Neural Network models and provided them with the prioritized importance of the prediction factors used in the study.
	Predicting Breast Cancer Survivability Using Data Mining Techniques	Bellaachia A. et al.	2006	The same study of Delen D. et al. but with newer data version and different data mining techniques: the Naïve Bayes, the back-propagated Neural Network, and the C4.5 Decision Tree algorithms using specificity and sensitivity as metrics of comparison. The results presented that C4.5 algorithm has a much better performance than the other two techniques
	Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction	Kumari M. et al.	2011	In critical field like medicine to predict cardiovascular disease patients, RIPPER, decision tree, Neural Networks and support vector machine prediction techniques performance were compared using the Sensitivity, Specificity, Accuracy, Error Rate, True Positive Rate and False Positive Rate as performance comparison metrics. The results showed that support vector machine model outperformed the other models for predicting cardiovascular disease.
	Comparison of Different Data Mining Techniques to Predict Hospital Length of Stay	Shailesh K R et. al.	2011	To predict the inpatient hospital length of stay in a super specialty hospital, the Neural Network was found to achieve better performance compared to the performance rates of Naive Bayes, K-NN, and decision tree prediction techniques.

Area of Research	Study Title	Author	Year	Main Contribution
Quality Control in Industry	A Comparison Between Data Mining Prediction Algorithms for Fault Detection (Case study: Ahanpishegan Co.)	Amooee G. et al.	2011	Prediction accuracy and processing time of the prediction techniques were the comparison performance metrics. The results showed that SVM and Logistic regression prediction algorithms has the best processing time with high overall prediction accuracy. The decision tree with its tree different branching algorithms (CRT, CHAID, and QUEST) achieved the highest prediction accuracy rates but needed more time. Neural Network achieved the least prediction accuracy rate with medium processing time.
	A Tree Based Model for High Performance Concrete Mix Design	C. Deepa et al.	2011	As performance metrics the prediction accuracy and error rates to predict the compressive strength of high performance concrete using MLP Neural Network, Rnd tree models and CRT regression was compared. The results showed that Neural Network and Rnd tree achieved the higher prediction accuracy rates and Rnd tree outperforms Neural Network regarding prediction error rates.
Education	A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year	Ho Yu C. et al.	2010	classification tree, multivariate adaptive regression splines (MARS), and Neural Network prediction performance were compared and the Neural Network was found to outperform the other two techniques. Also the results showed that transferred hours, residency, and ethnicity are crucial factors as independent variables to retention, which differs from previous studies.
	Predicting Students' Academic Performance, Comparing Artificial Neural Network, Decision Tree and Linear Regression	Ibrahim Z. et al.	2007	The Neural Network outperformed both the decision tree and linear regression models when the performance for the students' academic performance in the undergraduate degree program was measured by predicting the final cumulative grade point average (CGPA) of the students upon graduation. The correlation coefficient analysis was used to identify the relationship of the independent variables with the predictors.

In this research we expanded on the work of Lahiri R. [1] and performed a comparison on the same statistical and data mining techniques, viz., Logistic Regression, Neural Network and Decision Tree and to identify the effect of training data sampling method and sample size over the these prediction techniques using the Palestinian's household's expenditure and consumption data (PECS) produced by the Palestinian Central Bureau of Statistics (PCBS) [26]. In addition, we carried on the work suggested by Lahiri's as future work by selecting more precise predictors that significantly define and affect the output. We intended to select the predictors with the aid of domain experts and also perform correlation test to support our independent variable selection. The dependent variable values distribution, "0" and "1", effect will be examined on the overall prediction accuracy and on the individual prediction accuracy of "0" and "1".

Chapter Three

Background

In this section the data used in the research was discussed mentioning its source and characteristics. Also a detailed description of the prediction techniques were discussed and explored.

3.1 The PECS Data

"The Palestinian Central Bureau of Statistics, PCBS [26], annually conducted a household expenditure and consumption survey (PECS). The basic goal of this survey is to provide a necessary database for formulating national policies at various levels. This database explore the contribution of the household sector to the Gross National Product (GNP), determining the poverty degree, and providing weighted data that reflects the relative importance of the consumption items to be employed to determine the index for rates and prices of items and services. The PECS results are a fundamental cornerstone in the process of studying the nutritional status in the Palestinian territory. Another statistics are highly dependent on the PECS (Household Expenditure and Consumption) data like the calculation of price index and living conditions. The methodology of the survey is summarized as follows:

- The sample is stratified cluster systematic random sample with two stages, and 12 sub samples were used as one sub sample for each month.

- The duration of the survey is 12 months. The design of the survey took into consideration the seasonality in the consumption where it varies from one season to another like expenditure on fruit, vegetables and cloths.
- Each household was provided with a registration form (diary) where household would fill in daily expenditures. A female fieldworker would visit the household repeatedly 8-10 times to ensure registration of household's consumption in the diary according to the adopted procedures.
- The registration period for each household is restricted to one month. Households with longer registration periods than one month are given less variance in the expenditure and consumption pattern. One of the disadvantages to longer registration periods is that households would get bored or forget to fill in the specified form. The UN\ILO recommendations call for a registration period of three to four weeks. PCBS [26] selected a four week registration period to cover household's expenditures on goods and services that are repeated during the month.
- Different time references were adopted for the items of household's expenditure and consumption. The daily expenditure on food and transportation items was given a one-month reference period. Durable goods and educational fees are given 12-months reference period excluding personal transportation which is extended to the previous three years. Regarding income, a one month and one year reference periods were used.

Regarding the sampling and sampling frame, the target population consists of all Palestinian households who are residing habitually in the Palestinian Territory during 2009. The sampling frame consists of all enumeration areas which were enumerated in Census 2007; each numeration area consists of buildings and housing units with average of about 120 households in it. These enumeration areas are used as primary sampling units in

the first stage of the sampling selection. The estimated sample size for the Expenditure and Consumption survey 2009 was 4,584 household for the whole year in addition of 115 households over sample. Thus, the final sample size is 4,699 households. The non-response rate is estimated for the total sample is around 20%. The sample was designed in two staged stratified cluster sample. In the first stage a selection of systematic random sample of 191 enumeration areas was performed. In the second stage a selection of systematic random sample of 24 households from each enumeration area selected in the first stage was performed. In Jerusalem Governorate (J1), 13 enumeration areas were selected; then in the second phase, a group of households from each enumeration area were chosen using census-2007 method of delineation and enumeration. This method was adopted to ensure household response is to the maximum to comply with the percentage of non-response as set in the sample design. Finally, the enumeration areas were distributed to twelve months and the sample for each quarter covers sample strata (Governorate, locality type)". [13]

The main unit of research in the PECS data is the Palestinian household. The PECS data originally consisted of five tables, viz., IDENTIFICATION: contains household identification data, ROSTER: contains household characteristics, DWELLING: contains the dwelling characteristics and household living conditions, MAINGROUPS: contains household's monthly consumption and expenditure by main groups, and MONTHLY_INCOME: contains the monthly household's income. From these five tables we derived the data of this research of 40 columns in one table with 3,080 cleaned records for 2009 year and 3,757 cleaned records for 2010 year. (see Appendix 1 and Appendix 2)

3.2 Classification Techniques

3.2.1 Logistic Regression

It is a type of regression analysis used for predicting the outcome of a binary dependent variable which can take only two possible values like ("0" and "1") or ("yes" and "no")

based on related independent variables. It is used to identify the relationship, expressed as a probability that has only two values, between a dependent variable and one or more independent variables. Logistic regression attempts to find the occurrence probability of a "1" output using a linear function of the inputs as shown below:

$$f(z) = \frac{e^{z}}{1+e^{z}} = \frac{1}{1+e^{-z}}$$

Where:

- $z = (\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \cdots + \beta_k x_k)$ which is a measure of the total contribution of all the independent variables used in the model and is known as the logit.
- $x_1, x_2, x_3, \ldots, x_k$ are the factors (independent variables) that affect the probability.
- β_0 is the intercept which is the value of z when the value of all independent variables is zero.
- and $\beta_1, \beta_2, \beta_3, \ldots, \beta_k$ are the coefficients of the factors which measure the contribution of the factor in the probability.

Like probability, the domain of logistic regression function is $(-\infty,\infty)$ and the range is [0,1] (Fig. 3.1).

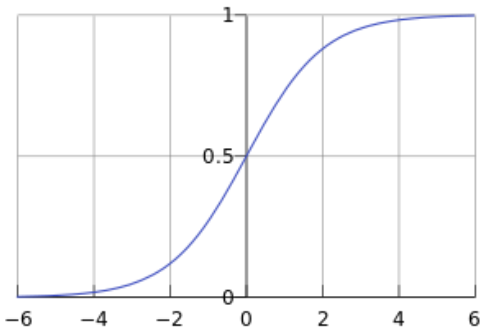

Figure 3.1: The logistic function, with z on the X axis and $f(z)$ on the Y axis [14].

An important use of logistic regression is for predicting binary outcomes and models a transformation of the expected value as a linear function of the predictors [14]. In this research we used logistic regression to predict the household's poverty status as it is seen in chapter four, the methodology section.

3.2.2 Decision Tree

One of the most popular classification and prediction techniques are the Decision Trees. They are easy to implement and understand by human because they are represented in the form of tree of nodes, where each node could be a *root*, father of all other nodes, that has no parent and include a test to be evaluated to split the tree into several sub-trees depending on a rule and the test results. Another node type is a *child* decision node, like the root but it has one father, containing further test to be implemented, establishing a sub-tree depending on the test results. The last type is a leaf node containing attribute value and no further tree splitting (Fig. 3.2).

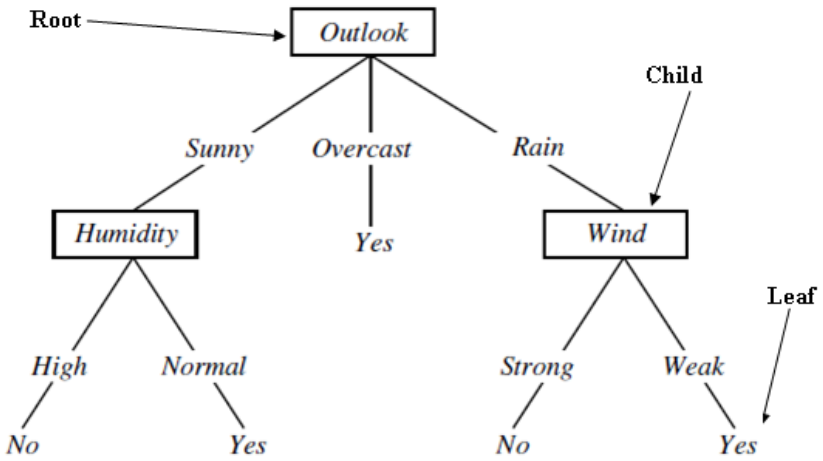

Figure 3.2: Decision Tree example of weather forecast [20].

Starting from the root and moving down in one route until to reach a leaf node is called a decision. So the decision tree is a set of decisions that classifies a set of data and provide a decision support mechanism (Fig. 3.3). To construct the tree, special classification algorithms are used, viz., CART, CHAID, C4.5, C5.0 and others. All these algorithms

create classification rules by constructing a tree-like structure of the data and they are different in the tree construction process trying to limit the size of the resulting tree [15]. In this research we built the prediction model using decision tree depending on CHAID algorithm (Chi-squared Automatic Interaction Detection). We tried also to use the CART algorithm and the results were the same, this is could be due to the data content and size, for more details see the methodology in chapter four.

Algorithm: Generate_decision_tree. Generate a decision tree from the training tuples of data partition *D*.

Input:

- Data partition, *D*, which is a set of training tuples and their associated class labels;
- *attribute_list*, the set of candidate attributes;
- *Attribute_selection_method*, a procedure to determine the splitting criterion that "best" partitions the data tuples into individual classes. This criterion consists of a *splitting_ attribute* and, possibly, either *a split point* or *splitting subset*.

Output: A decision tree.

Method:

(1) create a node *N*;

(2) **if** tuples in *D* are all of the same class, *C* **then**

(3) return *N* as a leaf node labeled with the class *C*;

(4) **if** *attribute_list* is empty **then**

(5) return *N* as a leaf node labeled with the majority class in *D*; // majority voting

(6) apply **Attribute_selection_method**(*D*, *attribute_list*) to find the "best" *splitting criterion*;

(7) label node *N* with *splitting_criterion*;

(8) i**f** *splitting_attribute* is discrete-valued **and**

multi-way splits allowed **then** // not restricted to binary trees

(9) *attribute_list* ← *attribute_list – splitting_attribute; //* remove *splitting_attribute*

(10) **for each** outcome *j* of *splitting_criterion*

// partition the tuples and grow sub-trees for each partition

(11) let D_j be the set of data tuples in *D* satisfying outcome *j*; // a partition

(12) **if** D_j is empty **then**

(13) attach a leaf labeled with the majority class in *D* to node *N*;

(14) **else** attach the node returned by **Generate_decision_tree**(D_j, *attribute_list*) to node *N*; **endfor**

(15) return *N*;

Figure 3.3: Basic algorithm for inducing a decision tree from training tuples [15].

3.2.3 Neural Network

Traditionally this term used to refer to a network of real biological neurons that are connected or functionally related in a nervous system. As inspiration of this biological system a mathematical or computational model was designed that simulate the structure and functional aspects of biological Neural Networks and called Artificial Neural Network (ANN) but when we use this term in information technology, we refer to it just by Neural Networks. Modern Neural Networks are non-linear data mining modeling tools used to model complex relationships between inputs and outputs or to recognize patterns in a given data set.

In the nervous system, a neuron collects signals from others through dendrites and sends out spikes of electrical activity through an axon, which splits into thousands of branches. At the end of each branch, a synapse converts the activity from the axon into electrical signals that prevent or activate activity in the connected neurons. If the activation input received by the neuron is larger than the prevention input, it sends a spike of electrical activity down its axon**.** The simulation of this real biological Neural Network, computational model, was programmed in computer and model learning occurs by benefitting from the knowledge of previous activities (Fig. 3.4 and Fig. 3.5).

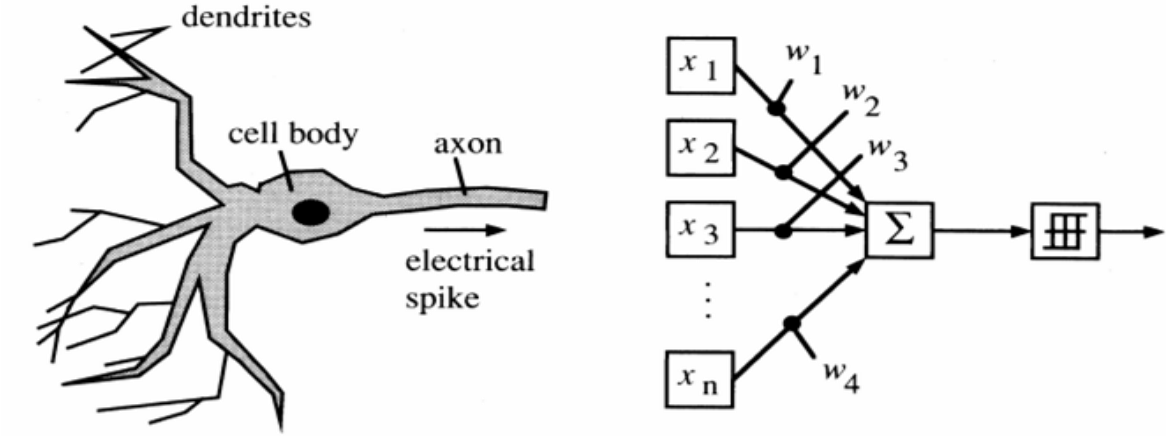

Figure 3.4: Biological Neural Network vs. Artificial Neural Network [21].

Algorithm: Backpropagation. Neural Network learning for classification or prediction, using the backpropagation algorithm.

Input:

- *D*, a data set consisting of the training tuples and their associated target values;
- *l*, the learning rate;
- *network*, a multilayer feed-forward network.

Output: A trained Neural Network.

Method:

(1) Initialize all weights and biases in *network*;

(2) **while** terminating condition is not satisfied {

(3) **for** each training tuple ***X*** in *D* f

(4) // Propagate the inputs forward:

(5) **for** each input layer unit *j* {

(6) $O_j = I_j$; // output of an input unit is its actual input value

(7) **for** each hidden or output layer unit *j* {

(8) $Ij = \sum_i w_{ij} O_i + \theta_j$; //compute the net input of unit *j* with respect to the previous layer, *i*

(9) $Oj = \frac{1}{1+e^{-Ij}}$;} // compute the output of each unit *j*

(10) // Backpropagate the errors:

(11) **for** each unit *j* in the output layer

(12) $Err_j = O_j(1-Oj)(T_j - O_j)$; // compute the error

(13) **for** each unit *j* in the hidden layers, from the last to the first hidden layer

(14) $Err_j = O_j(1-O_j) \sum_k Err_k w_{jk}$; // compute the error with respect to the next higher layer, *k*

(15) **for** each weight w_{ij} in *network* {

(16) $\Delta w_{ij} = (l) Err_j O_i$; // weight increment

(17) $w_{ij} = w_{ij} + \Delta w_{ij}$; } // weight update

(18) **for** each bias θ_j in *network* {

(19) $\Delta\theta_j = (l) Err_j$; // bias increment

(20) $\theta_j = \theta_j + \Delta\theta_j$; g // bias update

(21) }}

Figure 3.5: Neural Network Back-propagation algorithm [15].

In this research, as it seen in the methodology in chapter four, we used Neural Network to establish a prediction model because Neural Network is a powerful tool for pattern recognition. In training the network model, the model make use of the outputs that have inputs and recognize the pattern. When the network is used on data including patterns that hasn't associated output with the inputs, the network assigns the output that corresponds to a taught input pattern that is least different from the given input pattern.

The Neural Network simulation idea is that each neuron (node) has a certain number of inputs each holding incoming signal (instance) with a level of importance associated with each input called weight. The input value of a node is the sum of the weighted input values from its incoming inputs, if the sum passes a predefined threshold, and an activation function generates the node output value using the node input value and passes the node output to other nodes in the network. The set of input nodes are called the *input layer* while the set of output nodes are called the *output* layer, and in between there are another layer (one or two) called *hidden layer*. This is called multilayer perceptron (MLP) (Fig. 3.6).

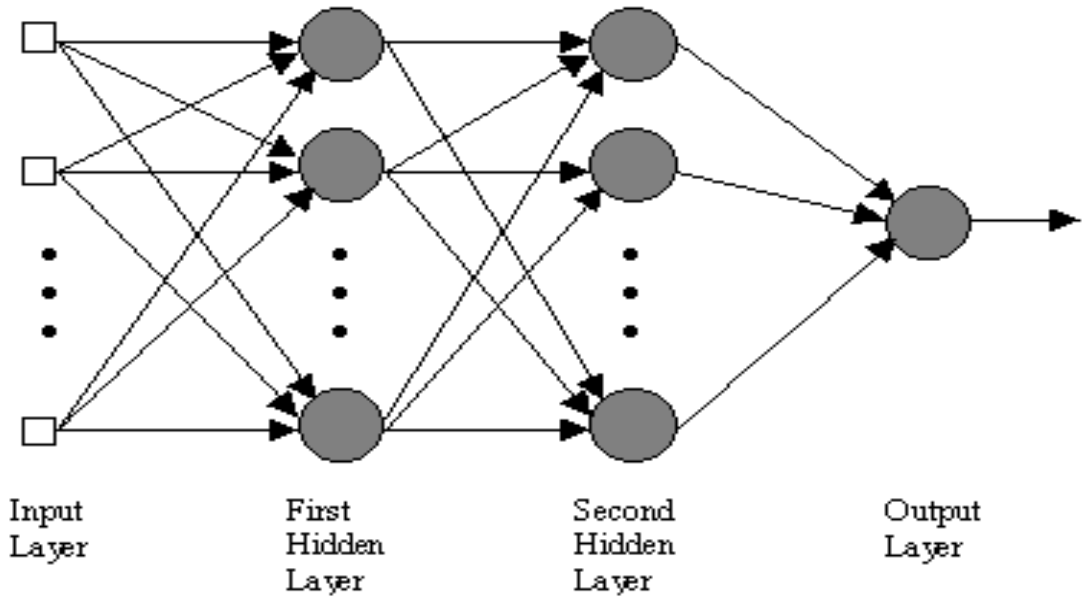

Figure 3.6: Neural Network Layers [22].

Each hidden unit is a function, the activation function, of the weighted sum of the inputs and the values of the weights are determined by the estimation algorithm. If the network contains a second hidden layer, each hidden unit in the second layer is a function of the weighted sum of the units in the first hidden layer. The activation function is a double sigmoid function as shown below:

$$f(\mathrm{sum_j}) = \frac{\mathrm{w_1}}{(1 + \exp(\mathrm{sum_j}))} + \frac{\mathrm{w_2}}{(1 + \exp(\mathrm{sum_j}))}$$

where sum_j is the scalar product of an input vector and weights to the node j either at a hidden layer or at the output layer and w1 and w2 are the initial weights. [15], [16], [17], [18], [19].

For this research, a multilayer perceptron (MLP) network and to specify the structure of the network the automatic architecture selection, which can select the best architecture automatically, was selected. For more details see the methodology in chapter four.

Chapter Four

Methodology

To achieve the objectives of this research, we started to prepare and clean the dataset, PECS 2009 and PECS 2010, in order to conduct the experiment. The original data were included in seven tables but the needed data for this research was contained in 5 tables, viz., IDENTIFICATION: contains household identification data, ROSTER: contains household characteristics, DWELLING: contains the dwelling characteristics and household living conditions, MAINGROUPS: contains household's monthly consumption and expenditure by main groups, and MONTHLY_INCOME: contains the monthly household's income. (see Appendix 1). From these five tables we derived the data for this research. It consists of 39 columns in one table with 3,080 cleaned records for 2009 year and 3,757 cleaned records for 2010 year (see and Appendix 2). The first step in preparing the data was to identify the dependent variable. With respect to the recommendations of this domain experts in PCBS [26], household's living conditions and social statistics, we identified the household's level of poverty as the dependent variable in this research. The household's level of poverty variable can be calculated depending on some information of the household exist in the aforementioned five tables but these information needed to be gathered in the same place. Therefore we worked on collecting and grouping the requested data from the five tables into one single data table depending on household identifier variable "ID00" that existed in all of the five tables which linked the household's data

together within the five tables. In addition to the household's data, some social and living conditions data can describe and affect the household's poverty status that comes from the head of household personal information also was added to the collected and aggregated data.

With the aid of this domain experts in PCBS [26] we had selected all variables from the separated data as independent variable that we believe they had a large effect and a big contribution in identifying the household into "poor" or "not poor", the values of the dependent variable. The resulted data table was cleaned against missing values by substituting the missing values with the exact value from other data sources for the same household. (see Appendix 2)

As a matter of fact, the dependent variable, level of poverty, was not actually existent in the data but it can be calculated depending on other variables and constraints, which what we had done. If the household's consumption value is less than the standard poverty line, determined by PCBS [26] for every year, then the household's poverty status is "poor" otherwise it is "not poor" and this holds for the standard household size, two parents and three children, while the households with different sizes, larger or smaller, has different poverty line which can be calculated using poverty line equivalence scale that used to assign the right poverty line value for the household depending on the household's size as follows:

- The equivalence scale for any household's size, denoted by *EqScale* is:

$$EqScale = (number\ of\ adults + 0.46 \times number\ of\ children)^{0.89}$$

- The equivalence scale for the reference household's size, 2 adult and 3 children:

$$EqScale_{Ref} = (2 + 0.46 \times 3)^{0.89} = 2.95621$$

- The poverty line for any household (h) is:

$$h = \left(\frac{EqScale_h}{EqScale_{Ref}}\right) \times Standard\ Poverty\ Line$$

- The standard poverty lines, identified by PCBS [26], for the years 2009 and 2010 were 2,168 New Israeli Shekels and 2,237 New Israeli Shekels respectively.

Depending on the principles and equations above we had calculated and assigned the right value of poverty line for each household. Thus we also assigned the actual level of poverty for each household, which it was the dependent variable, and assigned the value "1" as "poor" and the value "0" as "non poor".

As the classification and prediction methods in this research required the data to be binary, "0" and "1" only, most of the selected independent variables were binary having values only "0" or "1", e.g.: The household has *"Private Car"*. Some of the independent variables were categorical and the rest were continuous. All the categorical variables were re-coded and converted into binary by two ways. The first one was by grouping some of the variable's expected values into "1" and the rest values into "0", e.g.: *"The main material used in building outside walls of housing unit"* had seven expected values and after grouping it became one variable, *"Stone"*, with binary expected values. The second way was by splitting the variable into two or more variables each had expected values "0" or "1", e.g.: The variable *"Area"* had four expected values and it was spitted into four variables, *"AreaNorth", "AreaMid", "AreaSouth" and "AreaGza"* , each had binary expected values. The benefit of last way was to identify the values that have high contribution and effect on the probability of assigning the household's poverty status level. Some of the independent variables were derived and calculated from the existing data and leaved as continuous, e.g.: *"Household's Density"* = Household size/Total number of

rooms. The resulted list of independent variable count was around ninety variables and because the variables were binary the normality test will fail, so we applied *"Bivariate Correlations"* test, that can be used for binary data, to filter these independent variables and selected the correlated variables with *"Correlation Coefficient"* value of 10% and more, and value of *"Significance"* less than 0.05. The independent variables count, after correlation test, then reached 39 variables (Appendix 2). This holds for both 2009 data and 2010 data.

To prepare the prediction models using the three prediction techniques, we exploited the 2009 data as training dataset. As the one of the objectives of this research was to identify the effect of training data sample size and sampling method on the prediction performance, we used two different data sampling techniques, viz., *"Simple Random*" and *"Stratified Method*" sampling techniques. Using the *SPSS* tool, we divided the samples of training datasets into eight sizes, viz., 200, 400, 800, 1000, 1500, 2000, 2500 and 3000. It is straight forward to derive the samples using *"Random*" sampling, but to derive the samples using *"Stratified*" sampling method; we selected the variable *"Area*" as stratifying variable. The results of this process, drawing samples, were generating 16 training data files, eight using the *"Simple Random*" method and the other eight using the *"Stratified*" method, each file, in the eight groups, has different sample size.

After preparing the data and drawing the different sized training datasets using two different sampling methods, we trained the prediction models of logistic regression, decision tree (CHAID algorithm) and Neural Network (Multi-Layered Perceptron algorithm) using the 16 training datasets files and tested all the models using the PECS 2009 and the PECS 2010 data. The results were collected and recorded for further analysis and discussion.

The results analysis showed that the difference between prediction accuracy values for the dependent variable values “0” and “1” in all models, except in Neural Network, was not comparable. We believed this was due to the variation in distribution of the dependent variable values “0” and “1” in the data, and to check this we extended the study and included further analysis regarding the dependent variable values distribution. We did a values frequency check for the dependent variable in the PECS 2009 and PECS 2010 data and found that the ratio of dependent variable values, “1:0”, and distribution was more than 1:3 for both years. We applied the frequency check also on the 16 datasets and found that ratio of dependent variable values, “1:0”, distribution was also more than 1:3 for all the 16 training datasets. This is because the training datasets samples were drawn from the PECS 2009 data. Depending on the first results analysis, it is seen that the sample size didn’t have significant effect on the models performance and prediction accuracy and when increasing the sample size more than 800. Thus we produced a new revised version training dataset, only one dataset of size 800 records, by equating the number of records that have dependent variable value of both “0” and “1”. The idea was to keep the records that have the lower count of dependent variable distribution in the data which it was the value of poor “1”, fortunately its count was 721, then to draw a stratified sample from the remaining records that have the higher dependent variable distribution count which have the value not poor “0”, of equal size to the other value. Table 4.1 shows the distribution of “0” and “1” in the dependent variable and the ratio of this distribution in all our datasets.

Table 4.1: Distribution and ratio of the dependent variable values.

Dataset	**1: poor**	**0: Not poor**	**Total**	**Ratio**
Original Training dataset	According to sample size			Around 1:3
Revised Training dataset	721	719	1440	1:1
PECS 2009	721	2359	3080	Around 1:3
PECS 2010	876	2881	3757	Around 1:3

After preparing the new revised training dataset, we again trained the prediction models of logistic regression, decision tree (CHAID algorithm) and Neural Network (Multi-Layered Perceptron algorithm) using the new revised training dataset and tested all the models using the PECS 2009 and the PECS 2010 data. The results were collected and recorded for further analysis and discussion.

The PASW Statistics (SPSS Release 18.0.0) from IBM was used in all operations, modeling and to calculate all the aforementioned statistical and data mining techniques and methods.

Chapter Five

Results and Discussion

Applying the three modeling techniques (Logistic Regression, Decision Tree, and Neural Network) on the PECS 2009 and 2010 datasets resulted the following:

5.1 Logistic Regression Results

The logistic regression model was built using the PECS dataset of the year 2009 as training dataset for different sample sizes to predict the household's poverty status and classify it into *poor*: 1 or *not-poor*: 0 household. The results were grouped into two sets of tables, one for each sampling method.

Table (5.1) shows the summary of the results when the simple random sampling method was used to prepare the different data sample sizes from the 2009 dataset.

Table 5.1: Logistic Regression result on the training data (random sampling).

Sample Size	**% Agree**		
	0	**1**	**Overall**
200	93.5	61.7	86.0
400	94.2	55.6	85.5
800	92.3	52.6	82.9
1000	91.9	54.9	82.9
1500	93.5	49.7	83.2
2000	94.4	47.5	84.0
2500	93.2	46.0	82.2
3000	93.4	46.9	82.5

Table (5.2) and Table (5.3) present the summary of results for different sample sizes using simple random sampling for the years 2009 and 2010 respectively as testing datasets. The models were built for each sample size and applied to test the validity of prediction for the whole dataset.

Table 5.2: Logistic Regression result on year 2009 data (random sampling).

Sample Size	**% Agree**		
	0	**1**	**Overall**
200	87.0	38.7	75.7
400	89.5	48.1	79.8
800	91.3	50.6	81.8
1000	90.7	53.0	81.9
1500	93.0	47.6	82.4
2000	94.3	45.6	82.9
2500	93.0	47.0	82.2
3000	93.3	46.9	82.4

Table 5.3: Logistic Regression result on year 2010 data (random sampling).

Sample Size	**% Agree**		
	0	**1**	**Overall**
200	89.8	31.6	76.2
400	93.9	28.2	78.6
800	91.6	35.5	78.5
1000	95.5	27.1	79.5
1500	96.8	20.8	79.1
2000	94.8	27.7	79.1
2500	94.8	29.1	79.5
3000	94.8	29.0	79.5

Fig. (5.1), Fig. (5.2) and Fig. (5.3) plotting the overall % agreement against the sample sizes for the training dataset, testing dataset for the year 2009 and testing dataset for the year 2010 respectively. The results showed that the % agreement values for the dependent variable values “0” and “1” are highly abnormal within the training and testing datasets.

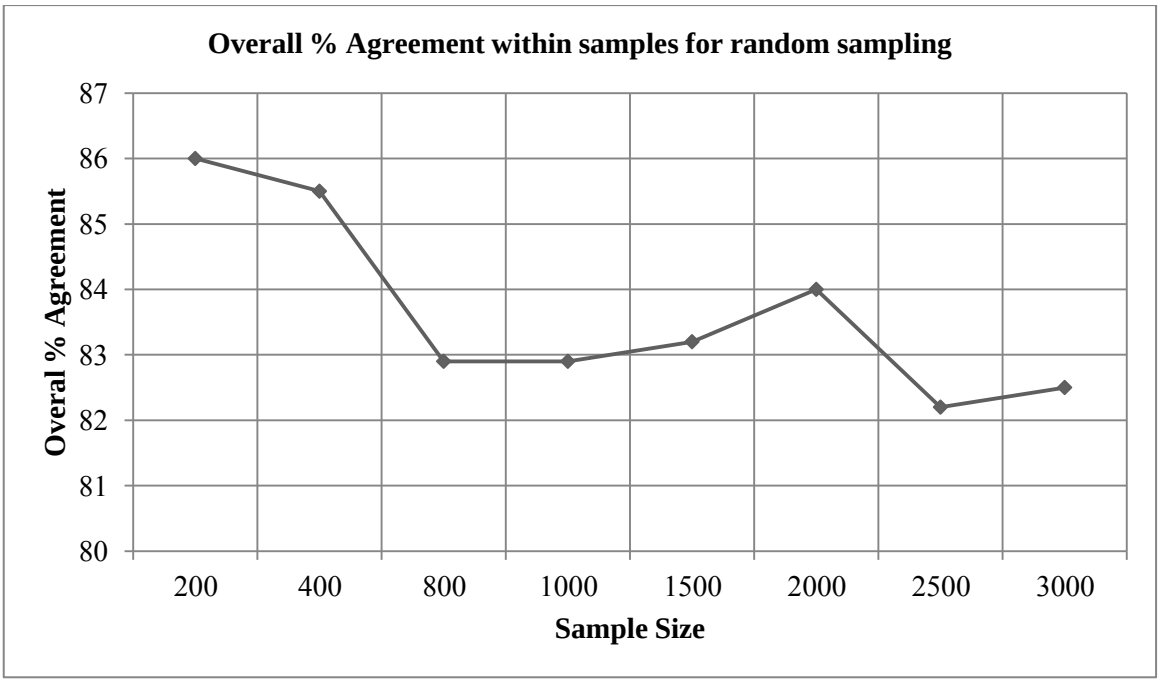

Figure 5.1: Logistic Regression result on the training data (random sampling).

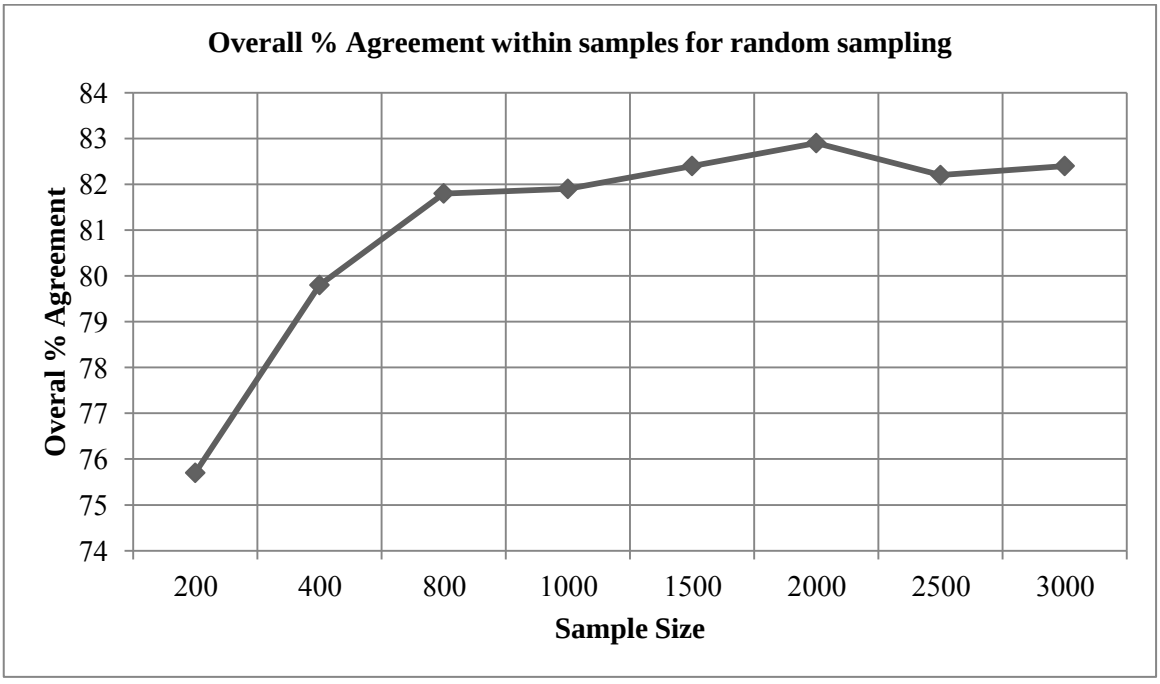

Figure 5.2: Logistic Regression result on year 2009 data (random sampling).

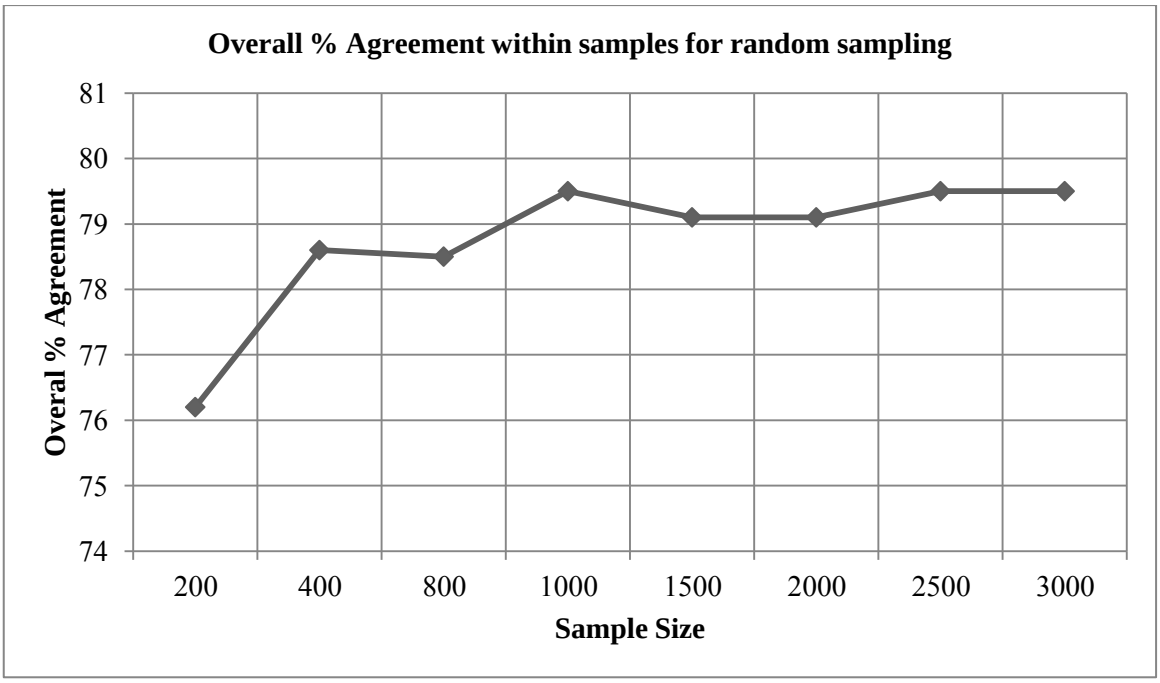

Figure 5.3: Logistic Regression result on year 2010 data (random sampling).

It is seen in Fig. (5.1) that the overall prediction accuracy for the training data was higher than that for the test data for the years 2009 and 2010 (Fig. 5.2 and Fig. 5.3), but it is closer to the 2009 testing data because the models were built using samples from year 2009 data. The overall prediction accuracy values for 2009 testing data mostly ranges in eighties like the values of the training data, all ranges in eighties, while all prediction accuracy values for 2010 testing data were ranges in seventies.

The results in Fig. (5.1) showed that the training dataset scored overall prediction accuracy of 86% when the sample size was 200 and it was the maximum overall prediction accuracy value. Then, by increasing the sample size the overall prediction accuracy dropped down and never increased again forming a decreasing curve. On the contrary, for the testing data and when the sample size was 200, the overall prediction accuracy started at minimum value of 75.7% and 76.2% for 2009 and 2010 data respectively. Then by increasing the sample size, the overall prediction accuracy increased forming an increasing curve.

For the training data, it is seen that a plateau was reached at sample size 800 where the overall prediction accuracy did not improve significantly when the sample size was increased beyond 800 which has an overall prediction accuracy of 82.9%. The testing results have the same behavior, it is seen that a plateau was reached also at the sample size of 800, where the overall prediction accuracy was 81.8% for 2009 data and 78.5% for 2010 data, and increasing the sample size beyond 800 did not significantly improve the overall prediction accuracy for the test data.

Alternatively, the logistic regression models were built again by using a stratified sampling method and using the household area as stratifying variable. Table (5.4) shows the summary of the results when the stratified sampling method was used to prepare the different data sample sizes from the training dataset.

Table 5.4: Logistic Regression result on the training data (stratified sampling).

Sample Size	**% Agree**		
	0	**1**	**Overall**
200	93.0	78.6	88.9
400	95.4	57.1	88.0
800	93.3	54.9	84.0
1000	93.6	59.2	85.2
1500	93.8	45.8	82.4
2000	92.7	50.2	83.0
2500	93.4	47.9	82.8
3000	93.0	47.7	82.3

Table (5.5) and Table (5.6) present the summary of results for different sample sizes using stratified sampling for the years 2009 and 2010 respectively as testing datasets. The models built for each sample size was applied to test the validity of prediction for the whole dataset.

Table 5.5: Logistic Regression result on year 2009 data (stratified sampling).

Sample Size	% Agree		
	0	1	Overall
200	81.0	61.7	76.5
400	93.1	42.2	81.2
800	92.1	48.1	81.8
1000	90.5	52.8	81.7
1500	93.2	45.5	82.0
2000	92.5	48.3	82.1
2500	92.8	47.9	82.3
3000	93.1	47.7	82.5

Table 5.6: Logistic Regression result on year 2010 data (stratified sampling).

Sample Size	% Agree		
	0	1	Overall
200	85.5	47.4	76.6
400	92.8	32.3	78.7
800	94.9	27.5	79.2
1000	91.8	38.1	79.3
1500	93.6	31.4	79.1
2000	93.5	32.5	79.3
2500	95.2	28.3	79.6
3000	95.0	28.8	79.5

Fig. (5.4), Fig. (5.5) and Fig. (5.6) plotting the overall % agreement against the sample sizes for the training dataset, testing dataset for year 2009 and testing dataset for the year 2010 respectively. The results showed that the prediction accuracy (% agreement) values for the dependent variable values “0” and “1” are highly abnormal within the training and testing data.

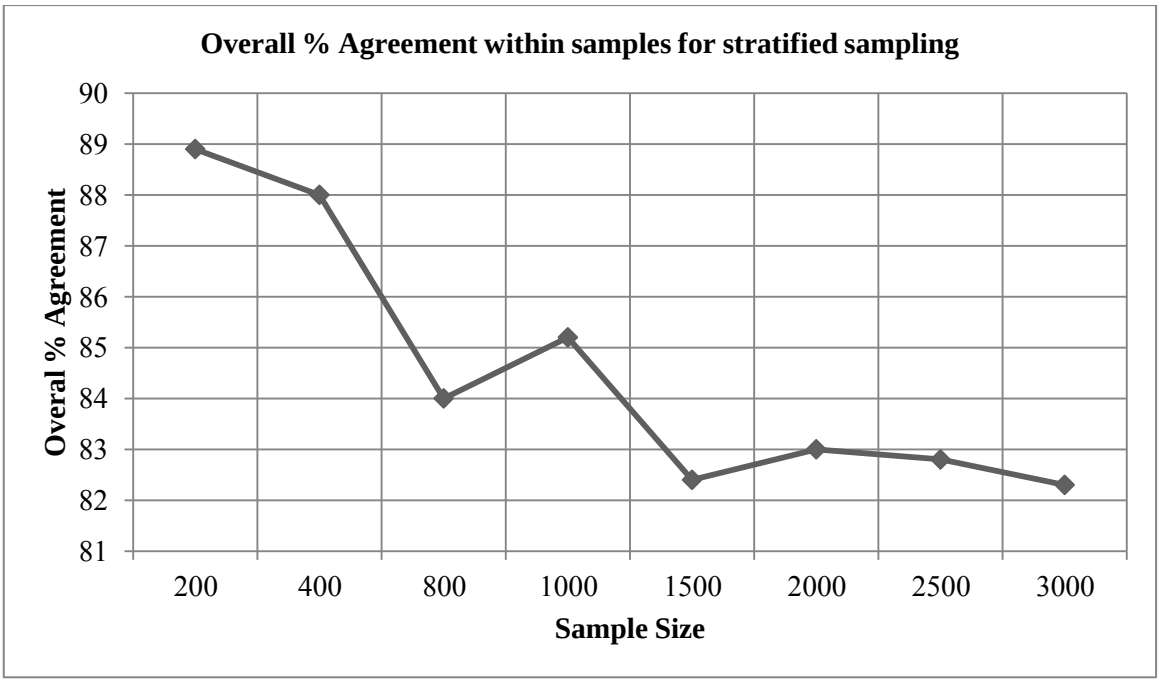

Figure 5.4: Logistic Regression result on the training data (stratified sampling).

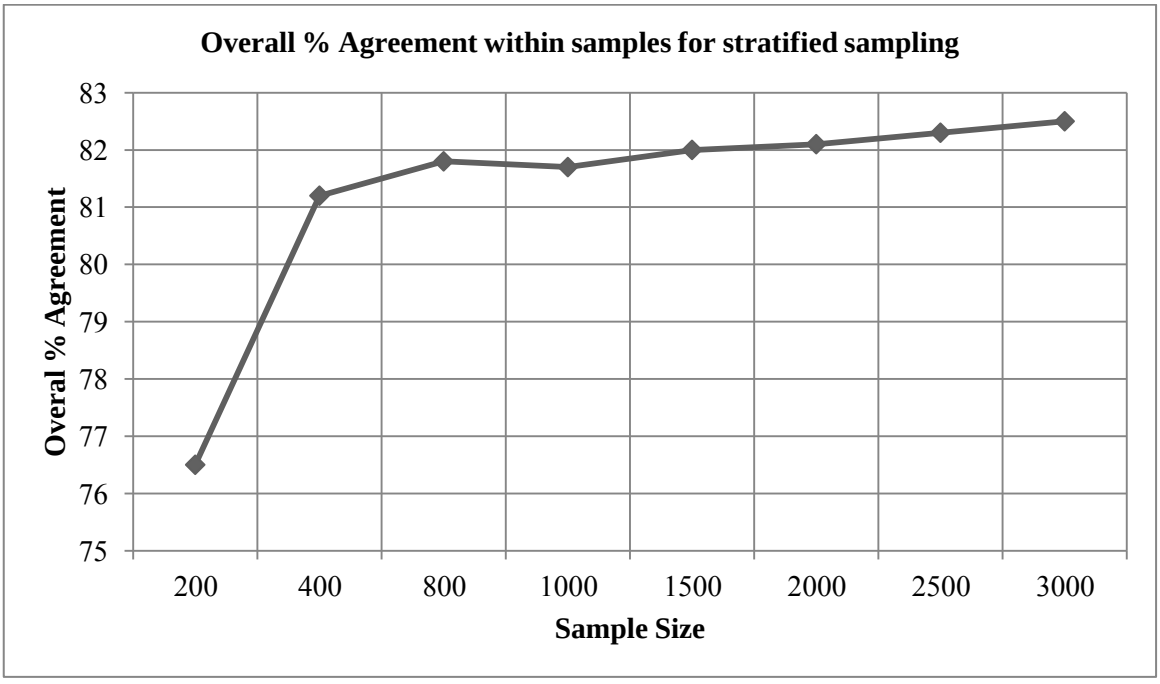

Figure 5.5: Logistic Regression result on year 2009 data (stratified sampling).

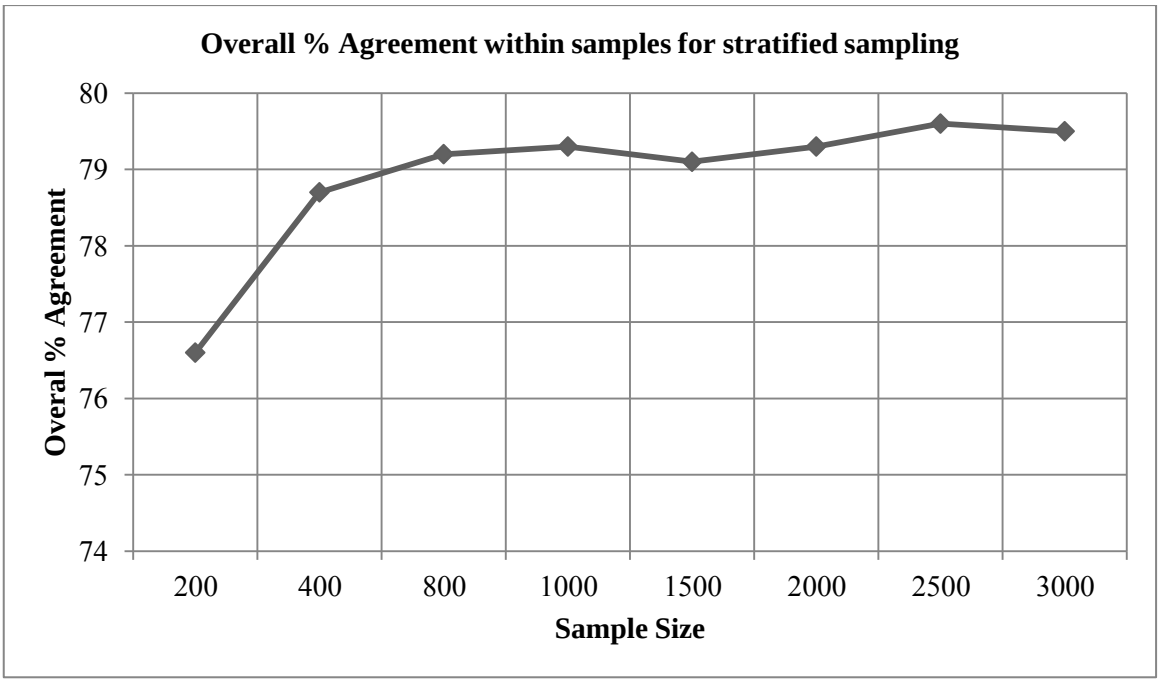

Figure 5.6: Logistic Regression result on year 2010 data (stratified sampling).

Applying logistic regression on different sizes sample datasets using stratified sampling method as seen in Fig. (5.4) that the overall prediction accuracy for the training data was higher than that for the test data for the year 2010 in Fig. (5.6) and almost the same for the test data for the year 2009 in Fig. (5.5), because the models were built using samples from year 2009 data. It is observed in Fig. (5.5) that the overall prediction accuracy values for 2009 data mostly ranges in eighties like the results of the training data in Fig. (5.4) that all ranges in eighties, while all prediction accuracy results for 2010 data in Fig. (5.6) are ranges in seventies.

The results in Fig. (5.4) showed that the training dataset scored overall prediction accuracy of 88.9% when the sample size was 200 and it was the maximum overall prediction accuracy value. Then, by increasing the sample size the overall prediction accuracy dropped down and never increased again forming a decreasing curve. On the contrary, for the testing data and when the sample size was 200, the overall prediction accuracy started at minimum value of 76.5% and 76.6% for 2009 data and 2010 data

respectively (Fig. 5.5 and Fig. 5.6), and by increasing the sample size the overall prediction accuracy increased forming an increasing curve.

As in the simple random sampling method results, for the training data in Fig. (5.4), it is seen that a plateau was reached at the sample size of 800 where the overall prediction accuracy did not improve when sample size was increased beyond 800 that has an overall prediction accuracy of 84%. The testing results have the same behavior, it is seen that in Fig. (5.5) and Fig. (5.6) a plateau was reached also at the sample size of 800, where the overall prediction accuracy was 81.8% for 2009 data and 79.2% for 2010 data and increasing the sample size beyond 800 did not significantly improve the overall prediction accuracy for the test data.

From the logistic regression results above, it is seen that the sampling method didn't have a significant influence on the prediction accuracy of the logistic regression technique with a very small outperformance for the stratified method.

5.2 Decision Tree Results

Like the logistic regression, the decision tree model was built using the PECS dataset of the year 2009 as training dataset for different sample sizes to predict the household's poverty status and classify it into poor: 1 or not-poor: 0 household. The results were grouped into two sets of tables, one for each sampling method.

Table (5.7) shows the summary of the results when the simple random sampling method was used to prepare the different data sample sizes from the 2009 dataset.

Table 5.7: Decision Tree result on the training data (random sampling).

Sample Size	% Agree		
	0	1	Overall
200	100.0	0.0	76.5
400	87.7	44.4	78.0
800	95.3	24.2	78.4
1000	96.2	26.6	79.2
1500	96.3	27.1	80.1
2000	96.3	30.2	81.6
2500	92.8	35.8	79.5
3000	96.3	26.5	80.0

Table (5.8) and Table (5.9) present the summary of results for different sample sizes using simple random sampling for the years 2009 and 2010 respectively as testing datasets. The models built for each sample size was applied to test the validity of prediction for the whole dataset.

Table 5.8: Decision Tree result on year 2009 data (random sampling).

Sample Size	% Agree		
	0	1	Overall
200	100.0	0.0	76.6
400	83.9	39.5	73.5
800	94.5	23.6	77.9
1000	95.5	26.2	79.3
1500	96.0	24.5	79.3
2000	96.0	27.2	79.9
2500	93.0	36.8	79.8
3000	96.3	26.1	79.9

Table 5.9: Decision Tree result on year 2010 data (random sampling).

Sample Size	% Agree		
	0	1	Overall
200	100.0	0.0	76.7
400	84.5	28.7	71.5
800	77.0	33.8	66.9
1000	95.3	18.4	77.3
1500	93.4	24.8	77.4
2000	96.9	12.4	77.2
2500	95.3	15.3	76.7
3000	95.9	20.3	78.3

Fig. (5.7), Fig. (5.8) and Fig. (5.9) plotting the overall % agreement against the sample sizes for the training dataset, testing dataset for year 2009 and testing dataset for the year

2010 respectively. The results in Fig. (5.7), Fig. (5.8) and Fig. (5.9) showed that there is a huge difference in the % agreement values for “0” and “1” values within the training and testing datasets.

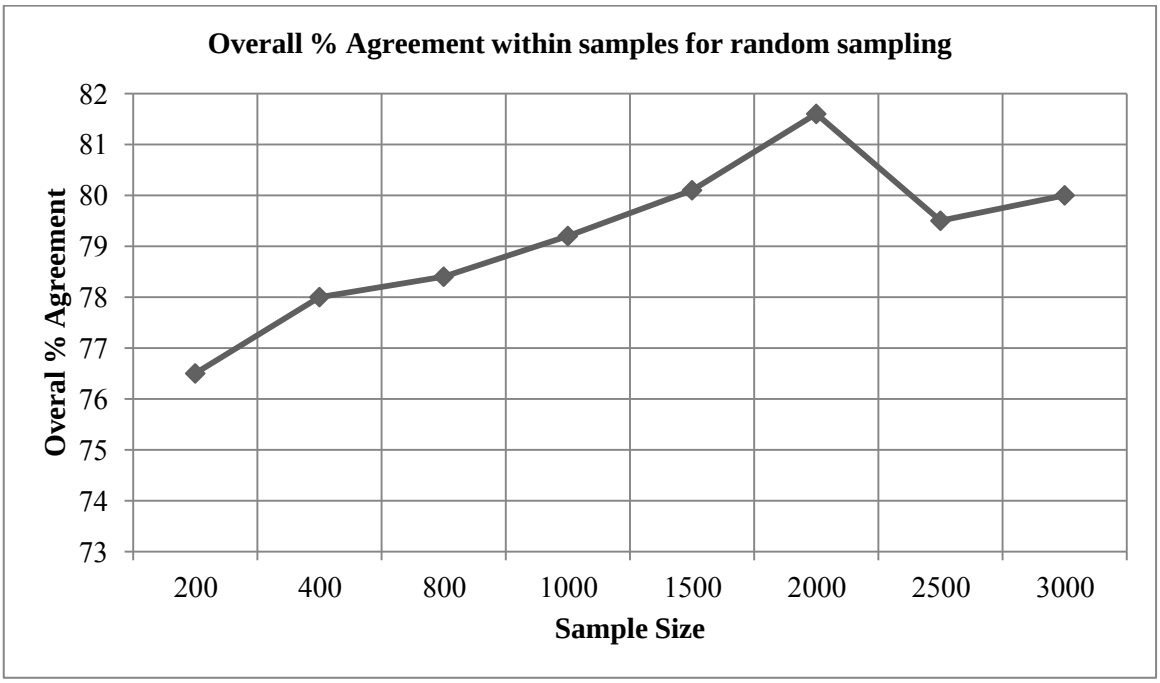

Figure 5.7: Decision Tree result on the training data (random sampling).

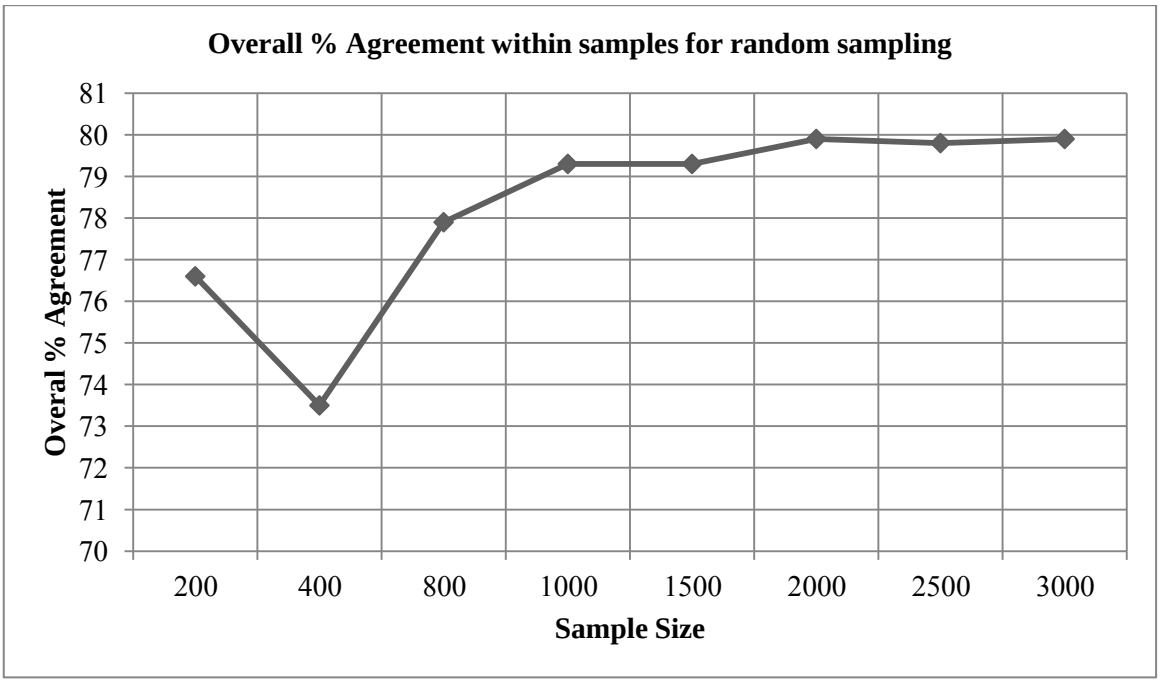

Figure 5.8: Decision Tree result on year 2009 data (random sampling).

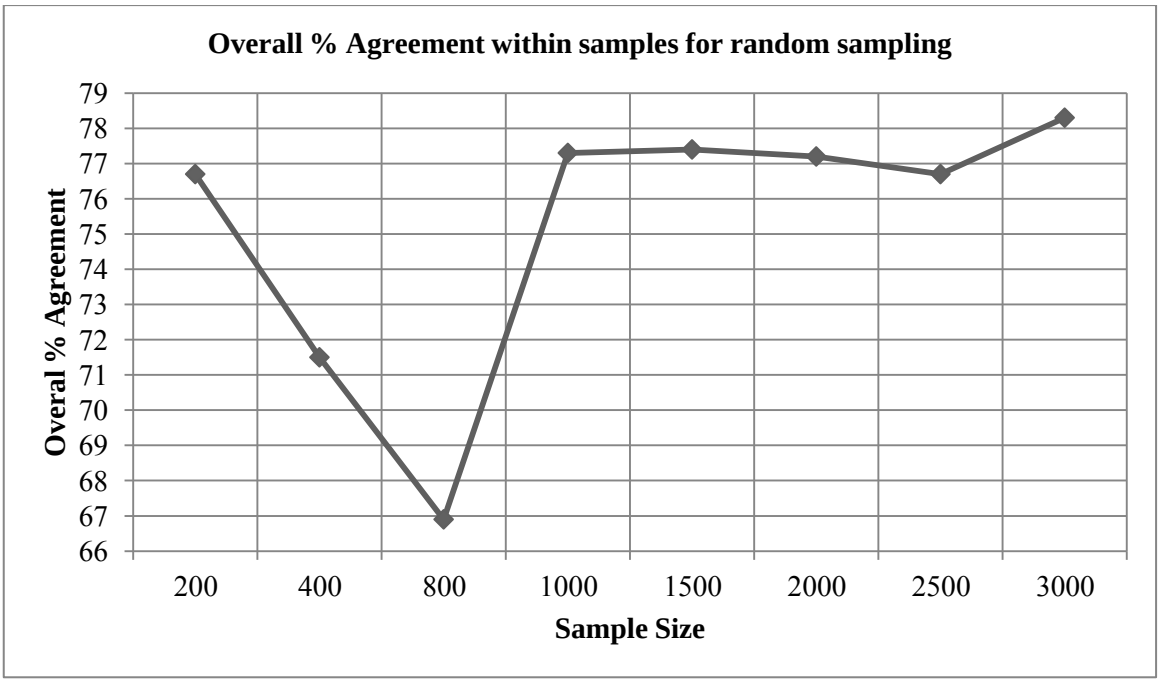

Figure 5.9: Decision Tree result on year 2010 data (random sampling).

It is seen in Fig. (5.7) that the overall prediction accuracy for the training data was higher than that for the test data for the year 2010 in Fig. (5.9) and almost the same of 2009 testing data in Fig. (5.8) because the models were built using samples from year 2009 data. It is seen that all the overall prediction accuracy values for 2009 data in Fig. (5.8) ranges at the end of seventies like the values of the training data in Fig. (5.7), most of them ranges at the end of seventies and three values are at the beginning of eighties, while all prediction accuracy values for 2010 data in Fig. (5.9) are ranges in seventies and one value was in sixties.

The results in Fig. (5.7) showed that the training dataset scored overall prediction accuracy of 81.6% when the sample size was 2000 and it was the maximum overall prediction accuracy value. Starting from sample size 200 and increasing the sample size, the overall prediction accuracy increased gradually forming an increasing curve. For the testing data in Fig. (5.8) and Fig. (5.9) the prediction accuracy values also plotted an increasing curves for both 2009 and 2010 data except when the sample size was 400 for

2009 testing data and 800 for 2010 testing data, the prediction accuracy dropped down to 73.5% and 66.9% respectively, then continued to rises up again.

For the training data in Fig. (5.7) it is seen that a plateau was reached at the sample size of 1000 where the overall prediction accuracy did not improve significantly when sample size was increased beyond 1000 that has an overall prediction accuracy of 79.2%, except of a small jump up when the sample size was 2000 the prediction accuracy reached a maximum value of 81.6% then it returns back to the plateau form. The testing results have the same behavior, it is seen that in Fig. (5.8) and Fig. (5.9) a plateau was reached also at the sample size of 1000, where the overall prediction accuracy was 79.3% for 2009 data and 77.3% for 2010 data, and increasing the sample size beyond 1000 did not significantly improve the overall prediction accuracy for the test data.

Decision tree models were built again by using a stratified sampling method and using the household area as stratifying variable. Table (5.10) shows the summary of the results when the stratified sampling method was used to prepare the different data sample sizes from the training dataset.

Table 5.10: Decision Tree result on the training data (stratified sampling).

Sample Size	**% Agree**		
	0	**1**	**Overall**
200	81.1	60.7	75.4
400	100.0	0.0	80.8
800	96.4	29.5	80.3
1000	98.3	20.8	79.3
1500	94.8	26.8	78.7
2000	96.2	29.3	80.9
2500	95.5	27.7	79.7
3000	96.3	25.8	79.7

Table (5.11) and Table (5.12) present the summary of results for different sample sizes testing datasets using stratified sampling for the years 2009 and 2010 respectively.

The models built for each sample size was applied to test the validity of prediction for the whole dataset.

Table 5.11: Decision Tree result on year 2009 data (stratified sampling).

Sample Size	**% Agree**		
	0	**1**	**Overall**
200	76.0	52.0	70.4
400	100.0	0.0	76.6
800	95.0	24.8	78.5
1000	97.2	20.4	79.2
1500	95.4	27.3	79.4
2000	96.3	27.3	80.2
2500	95.4	27.3	79.4
3000	96.3	26.1	79.9

Table 5.12: Decision Tree result on year 2010 data (stratified sampling).

Sample Size	**% Agree**		
	0	**1**	**Overall**
200	74.2	50.2	68.6
400	100.0	0.0	76.7
800	95.8	16.3	77.3
1000	99.9	0.8	76.8
1500	91.5	25.5	76.1
2000	96.6	14.8	77.5
2500	91.5	25.5	76.1
3000	95.9	20.3	78.3

Fig. (5.10), Fig. (5.11) and Fig. (5.12) plotting the overall % agreement against the sample sizes for the training dataset, testing dataset for year 2009 and testing dataset for the year 2010 respectively. The results in Fig. (5.10), Fig. (5.11) and Fig. (5.12) showed that there is a huge difference in the % agreement values for “0” and “1” values within the training and testing datasets.

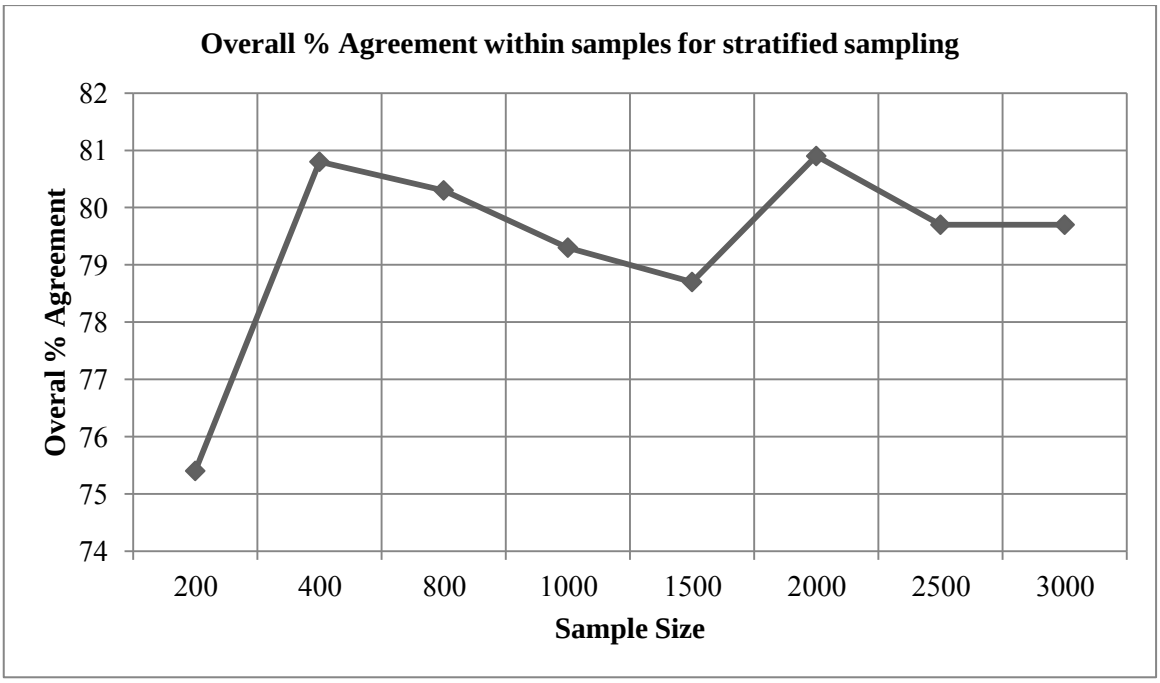

Figure 5.10: Decision Tree result on the training data (stratified sampling).

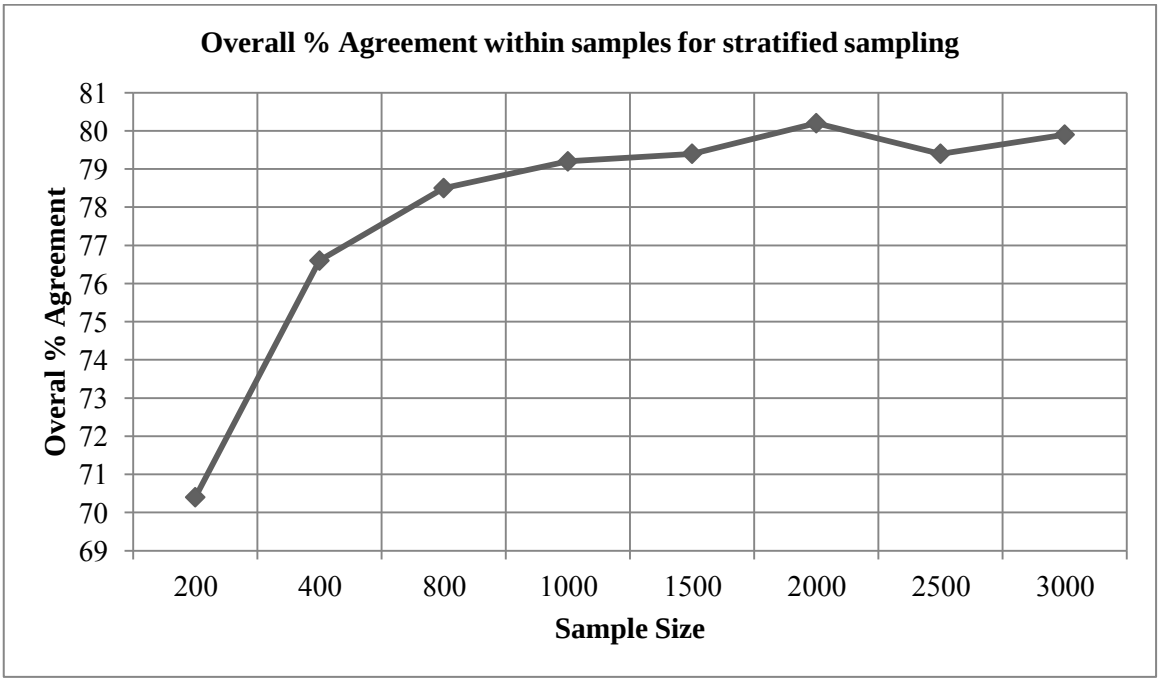

Figure 5.11: Decision Tree result on year 2009 data (stratified sampling).

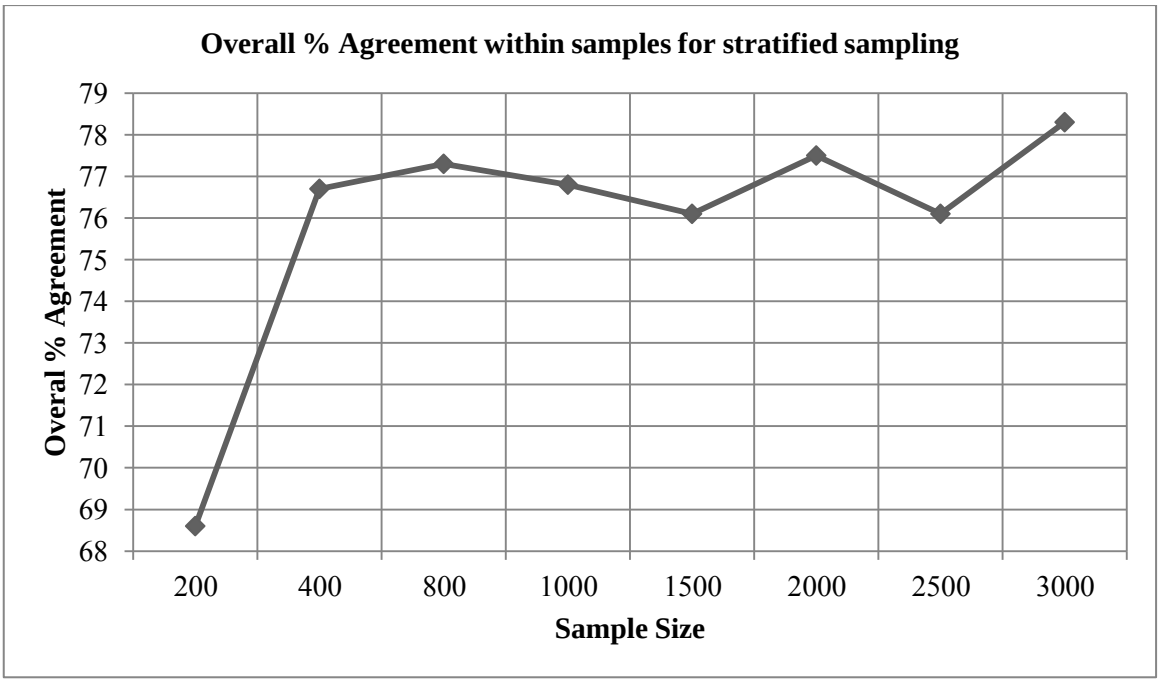

Figure 5.12: Decision Tree result on year 2010 data (stratified sampling).

The results in Fig. (5.10) showed that the training dataset plot two-humped curve and reached maximum overall prediction accuracy of 80.9% and then faltered off. It is seen that the overall prediction accuracy for the training data in Fig. (5.10) was slightly higher than that for the test data for the year 2010 data in Fig. (5.12) and almost the same of 2009 testing data in Fig. (5.11) because the models were built using samples from year 2009 data. It is seen in Fig. (5.10), Fig. (5.11) and Fig. (5.12) that most of the overall prediction accuracy values for the training data and the testing data, 2009 and 2010, ranges at the ends of seventies except five values are not in this range. Three overall prediction accuracy values in the training data and one value in the 2009 testing data are around 80%, while one overall prediction accuracy value in the 2010 data is around 68%.

The results in Fig. (5.10) showed that the training dataset scored overall prediction accuracy around 80% and it was the maximum overall prediction accuracy value. Starting from sample size 200 and increasing the sample size, the overall prediction accuracy increased slightly at some parts and decreased slightly at the other parts of the graph. In the

testing data the prediction accuracy values plotted an increasing curves for both 2009 data in Fig. (5.11) and 2010 data in Fig. (5.12) except when the sample size was 1500 for 2010 testing data, the prediction accuracy dropped down to 76.1%, then continued to rises up again.

For the training data in Fig. (5.10), it is seen that a plateau was reached at the sample size of 800 where the overall prediction accuracy did not improve significantly when sample size was increased beyond 800 that has an overall prediction accuracy of 80.3%, except of a small jump down when the sample size was 1500 where the prediction accuracy reached a value of 78.7% then it returns back to the rise up to the plateau form. It is seen that the testing data of both 2009 data in Fig. (5.11) and 2010 data in Fig. (5.12) reached a plateau at the sample size of 800, where the overall prediction accuracy was 78.5% for 2009 data and 77.3% for 2010 data, and increasing the sample size beyond 800 did not significantly improve the overall prediction accuracy for the test data.

From the decision tree results above, it is seen that the sampling method didn’t have a significant influence on the prediction accuracy of the decision tree technique.

5.3 Neural Network Results

The Neural Network model was built using the PECS dataset of the year 2009 as training dataset for different sample sizes to predict the household's poverty status and classify it into poor: 1 or not-poor: 0 household. The results were grouped into two sets of tables, one for each sampling method.

Table (5.13) shows the summary of the results when the simple random sampling method was used to prepare the different data sample sizes from the 2009 dataset.

Table 5.13: Neural Network result on the training data (random sampling).

Sample Size	% Agree		
	0	1	Overall
200	96.7	76.6	92.0
400	98.7	64.4	91.0
800	96.6	85.8	94.0
1000	95.0	85.2	92.6
1500	95.8	71.5	90.1
2000	96.5	67.6	90.1
2500	93.2	64.8	86.6
3000	94.0	60.9	86.2

Table (5.14) and Table (5.15) present the summary of results for different sample sizes using simple random sampling for the years 2009 and 2010 respectively as testing datasets. The models built for each sample size was applied to test the validity of prediction for the whole dataset.

Table 5.14: Decision Tree result on year 2009 data (random sampling).

Sample Size	% Agree		
	0	1	Overall
200	88.7	45.0	78.5
400	85.5	57.9	79.0
800	90.7	60.3	83.6
1000	87.2	65.9	82.2
1500	92.6	59.8	84.9
2000	94.3	58.5	85.9
2500	92.7	62.3	85.6
3000	93.3	66.7	87.0

Table 5.15: Neural Network result on year 2010 data (random sampling).

Sample Size	% Agree		
	0	1	Overall
200	86.6	46.1	77.2
400	97.2	16.8	78.4
800	90.4	37.8	78.1
1000	93.9	28.3	78.6
1500	93.5	29.0	78.4
2000	88.7	37.7	76.8
2500	92.5	35.3	79.2
3000	92.5	35.3	79.2

Fig. (5.13), Fig. (5.14) and Fig. (5.15) plotting the overall % agreement against the different sample sizes for the training dataset, testing dataset for year 2009 and testing dataset for the year 2010 respectively. The results in Fig. (5.13), Fig. (5.14) and Fig. (5.15) showed that the prediction accuracy (% agreement) values for "0" and "1" values are comparable for the training data and nearly for 2009 testing data while it was abnormal for 2010 testing data.

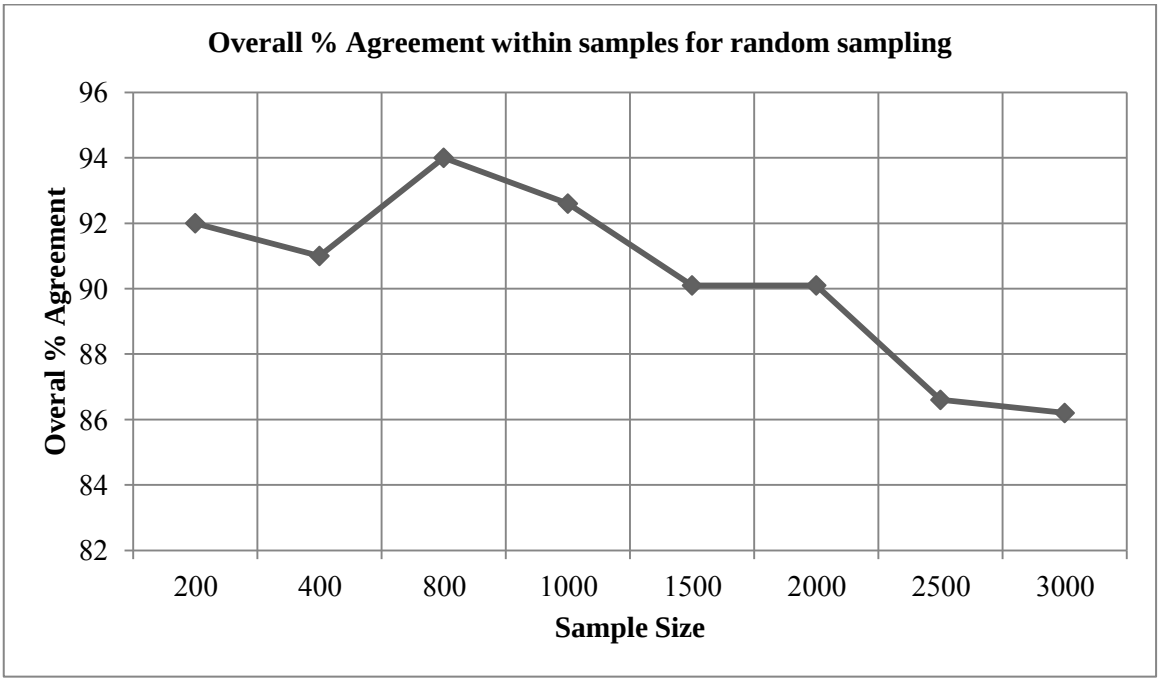

Figure 5.13: Neural Network result on the training data (random sampling).

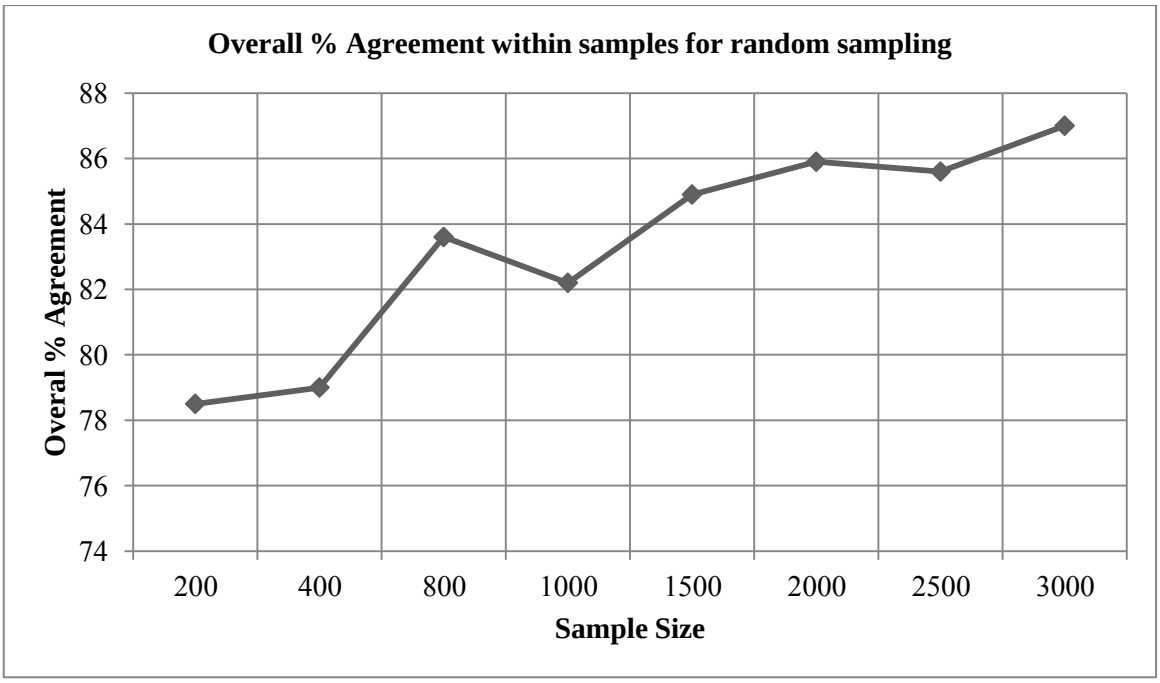

Figure 5.14: Neural Network result on year 2009 data (random sampling).

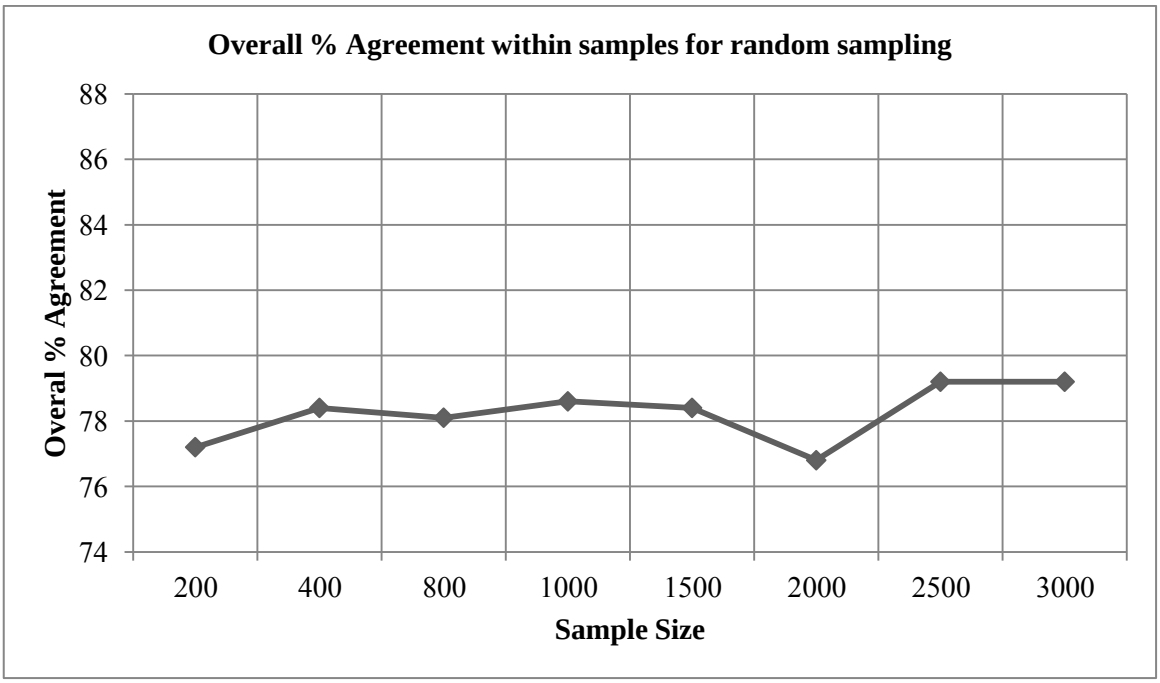

Figure 5.15: Neural Network result on year 2010 data (random sampling).

The Neural Network results above showed that the overall prediction accuracy for the training data in Fig. (5.13) was higher than that for the test data for the years 2009 in Fig. (5.14) and 2010 in Fig. (5.15) but it is closer to the 2009 testing data because the models were built using samples from year 2009 data. It is seen that the overall prediction

accuracy of the training data in Fig. (5.13) reached maximum value when the sample size is 800 with prediction accuracy around 94% then it keeps dropping down forming a decreasing curve. On the contrary, the overall prediction accuracy for both 2009 in Fig. (5.14) and 2010 testing data in Fig. (5.15) plotted an increasing curves except when the sample size was 2000 for 2010 testing data where the prediction accuracy dropped down to 76.8% then continued to rises up again. Most of the overall prediction accuracy values for training data in Fig. (5.13) ranges in nineties, while for 2009 testing data in Fig. (5.14) most of the values ranges in eighties, and in 2010 testing data in Fig. (5.15) all values ranges in seventies.

For the training data in Fig. (5.13), it is seen that no plateau was plotted and the overall prediction accuracy values changed significantly up and down for different sample sizes. Also in the testing data in Fig. (5.14) and Fig. (5.15) there are no clear plateau plotted but partial plateau reached when the sample size is 800 for both 2009 and 2010 then the overall prediction accuracy values continued to rise significantly with the increase of sample size except when the sample size is 2000 in the 2010 testing where the overall prediction accuracy reached minimum value of 76.8% then continued to rises up again.

Neural Network models were also built again by using a stratified sampling method and using the household area as stratifying variable. Table (5.16) shows the summary of the results when the stratified sampling method was used to prepare the different data sample sizes from the training dataset.

Table 5.16: Neural Network result on the training data (stratified sampling).

Sample Size	% Agree		
	0	1	Overall
200	94.4	83.9	91.5
400	94.1	88.3	93.0
800	97.7	79.8	93.4
1000	97.0	84.1	93.8
1500	96.3	70.9	90.3
2000	97.4	81.2	93.7
2500	94.8	66.6	88.2
3000	94.2	63.3	86.9

Table (5.17) and Table (5.18) present the summary of results for different sample sizes testing datasets using stratified sampling for the years 2009 and 2010 respectively. The models built for each sample size was applied to test the validity of prediction for the whole dataset.

Table 5.17: Neural Network result on year 2009 data (stratified sampling).

Sample Size	% Agree		
	0	1	Overall
200	84.4	61.8	79.1
400	87.4	61.2	81.3
800	91.3	55.0	82.8
1000	88.3	65.6	83.0
1500	91.7	59.8	84.2
2000	93.1	60.7	85.5
2500	93.6	64.1	86.7
3000	93.4	67.7	87.4

Table 5.18: Neural Network result on year 2010 data (stratified sampling).

Sample Size	% Agree		
	0	1	Overall
200	86.1	48.9	77.4
400	89.4	39.0	77.6
800	91.6	33.7	78.1
1000	86.8	42.9	76.6
1500	87.6	40.3	76.6
2000	88.0	39.9	76.8
2500	92.0	34.7	78.7
3000	94.0	30.6	79.2

Fig. (5.16), Fig. (5.17) and Fig. (5.18) plotting the overall % agreement against the sample sizes for the training dataset, testing dataset for year 2009 and testing dataset for the year 2010 respectively. The results in Fig. (5.16), Fig. (5.17) and Fig. (5.18) showed that the prediction accuracy (% agreement) values for "0" and "1" values are comparable for the training data in Fig. (5.16) and nearly for 2009 testing data in Fig. (5.17) while it was abnormal for 2010 testing data in Fig. (5.18).

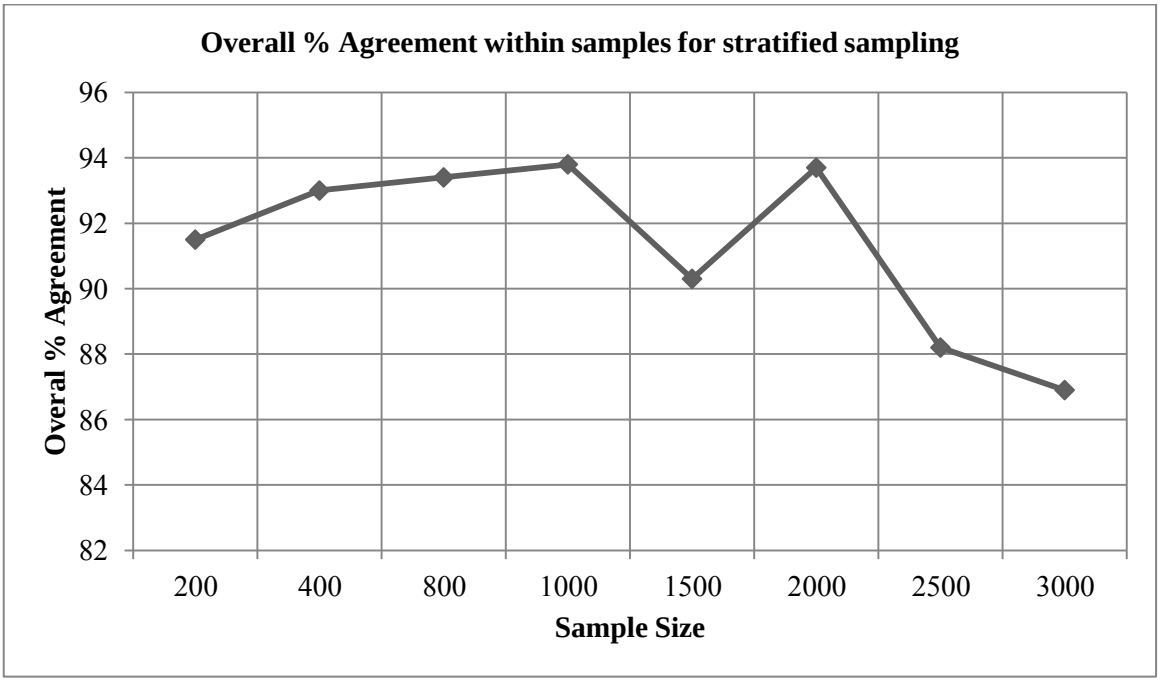

Figure 5.16: Neural Network result on the training data (stratified sampling).

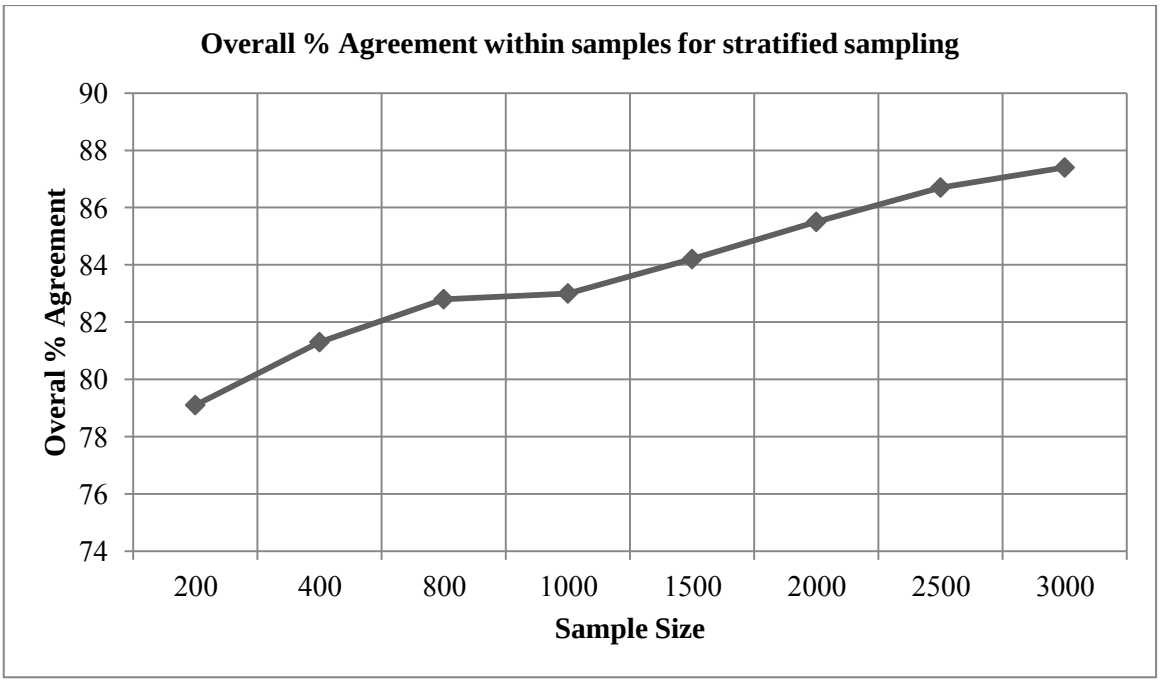

Figure 5.17: Neural Network result on year 2009 data (stratified sampling).

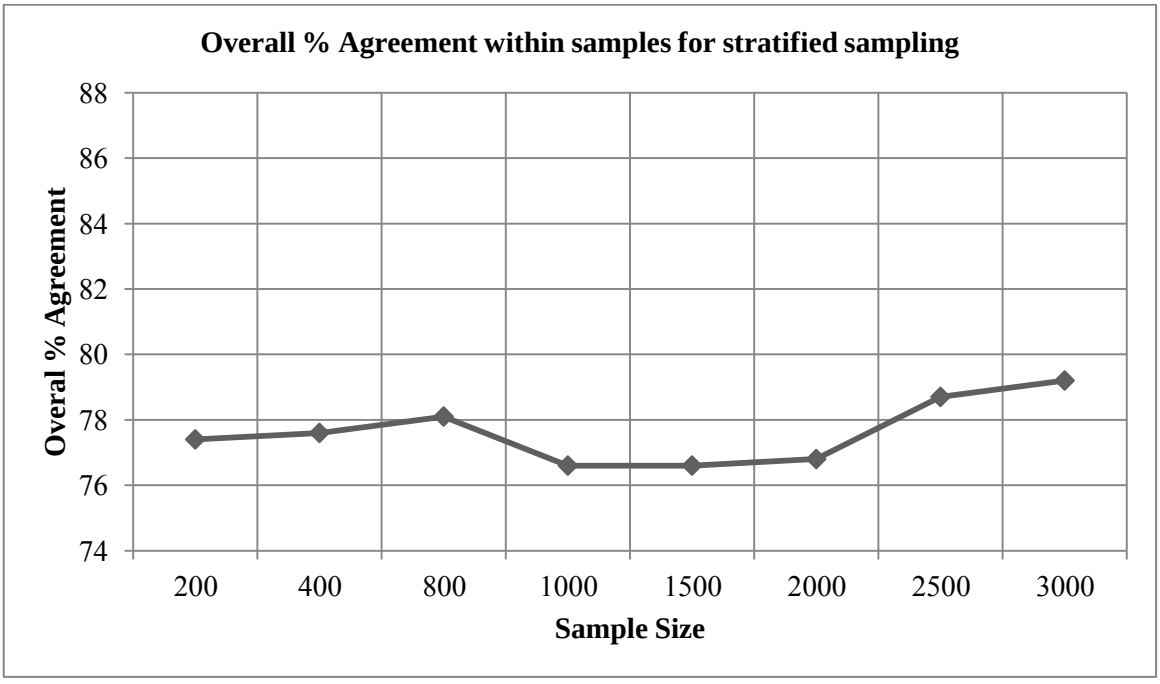

Figure 5.18: Neural Network result on year 2010 data (stratified sampling).

The Neural Network results above showed that the overall prediction accuracy for the training data in Fig. (5.16) was higher than that for the testing data for the years 2009 in Fig. (5.17) and 2010 data in Fig. (5.18) but it is closer to the 2009 testing data because the

models were built using samples from year 2009 data. It is seen that the overall prediction accuracy of the training data in Fig. (5.16) increased slightly with the increase of sample size forming an increasing curve until it reaches sample size 1500 it dropped down to around 90% and it increased again at sample size 2000 then finally keep dropping down for the remaining sample sizes. Most of the overall prediction accuracy values for the training data ranges in nineties. The results showed that the overall prediction accuracy values for both 2009 data in Fig. (5.17) and 2010 data in Fig. (5.18) plotted an increasing curves except when the sample size was 1000 for 2010 testing data where the prediction accuracy dropped down to 76.6%, then continued to rises up again. Most of the overall prediction accuracy values for 2009 testing data ranges in eighties, and in 2010 testing data all values ranges in seventies.

For the training data in Fig. (5.16) an increasing curve was plotted until it reached maximum value of 93.8% at sample size 1000 then the curve was disturbed up and down. A partial plateau was plotted between the sample size 400 to 1000 and the overall prediction accuracy values changed significantly up and down for different remaining sample sizes. For the 2009 testing data in Fig. (5.17) an increasing curve was plotted and a small plateau observed between sample size 800 and 100, but the curve continued rising up significantly. Also for 2010 testing data in Fig. (5.18) an increasing curve was plotted but dropped down at sample size 800 forming partial plateau between sample size 1000 to 2000 then continued to rise up significantly.

From the Neural Network results above, it is seen that the sampling method didn't have a significant influence on the prediction accuracy of the decision tree technique. It is noticed that in the training data, when increasing sample size up to 800, the overall prediction accuracy rises significantly and increasing the sample size more than 800 it

decreased the overall prediction accuracy. This is in a contrary with testing data because the overall prediction accuracy mostly increased every time the sample size increased.

To examine the effect of the dependent variable's values distribution on the prediction accuracy of our prediction techniques, an additional experiment was conducted, as mentioned in the methodology in chapter 4. The results of the additional experiment was discussed in the next section.

5.4 Revised Training Data Results

When the additional analysis was conducted to improve the prediction accuracy depending on equating the dependent variable's values distribution, "0" and "1", the three models, (logistic regression, decision tree, and Neural Network), was built again and trained by using the new revised version dataset and tested by the whole dataset of 2009 and 2010 PECS data. Table (5.19) shows the summary of results of prediction accuracy after the prediction models were rebuilt and trained using the new revised training data, then tested by using the whole datasets of PECS 2009 and 2010.

Table 5.19: Prediction accuracy results of the new revised training data.

Model		**% Agree**		
		0	**1**	**Overall**
Logistic Regression	Training	75.8	81.0	78.4
	Test 2009	73.5	81.0	75.3
	Test 2010	78.0	68.5	75.8
Decision Tree	Training	70.2	75.6	72.9
	Test 2009	68.0	75.6	69.8
	Test 2010	81.2	41.3	71.9
Neural Network	Training	87.9	92.1	90.0
	Test 2009	76.1	92.1	79.8
	Test 2010	78.9	66.4	76

Fig. (5.19) and Fig. (5.20) plot the overall prediction accuracy of the models using each training data, 2009 testing data and 2010 testing data.

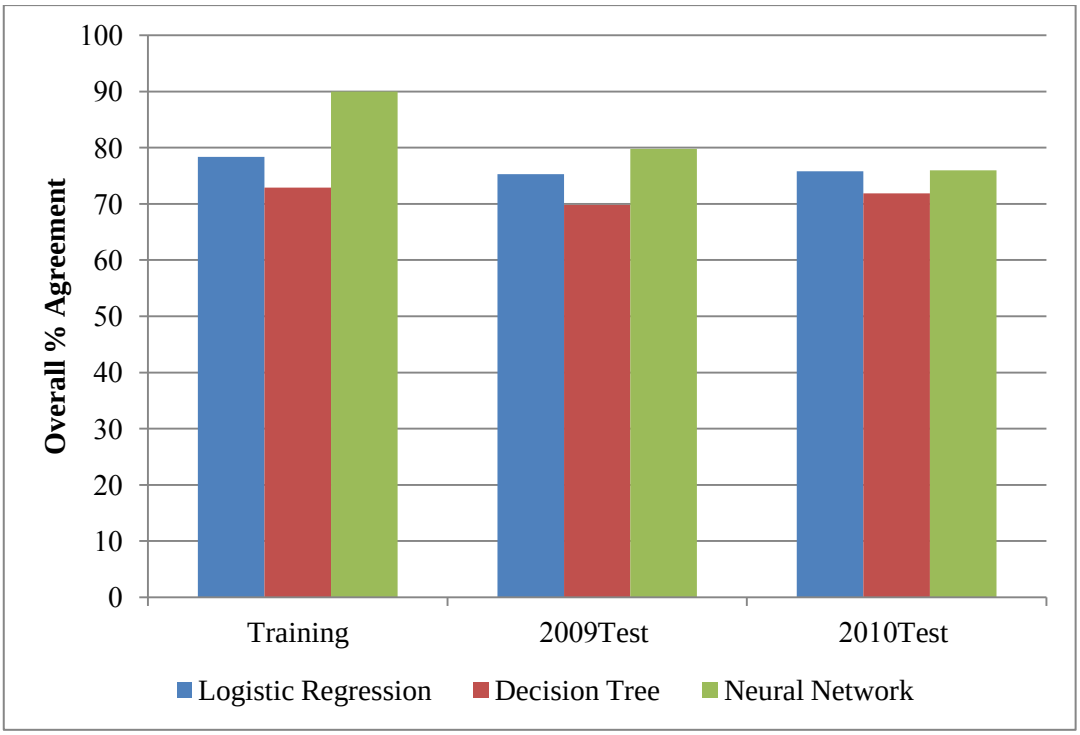

Figure 5.19: Overall prediction accuracy of prediction models in the revised analysis.

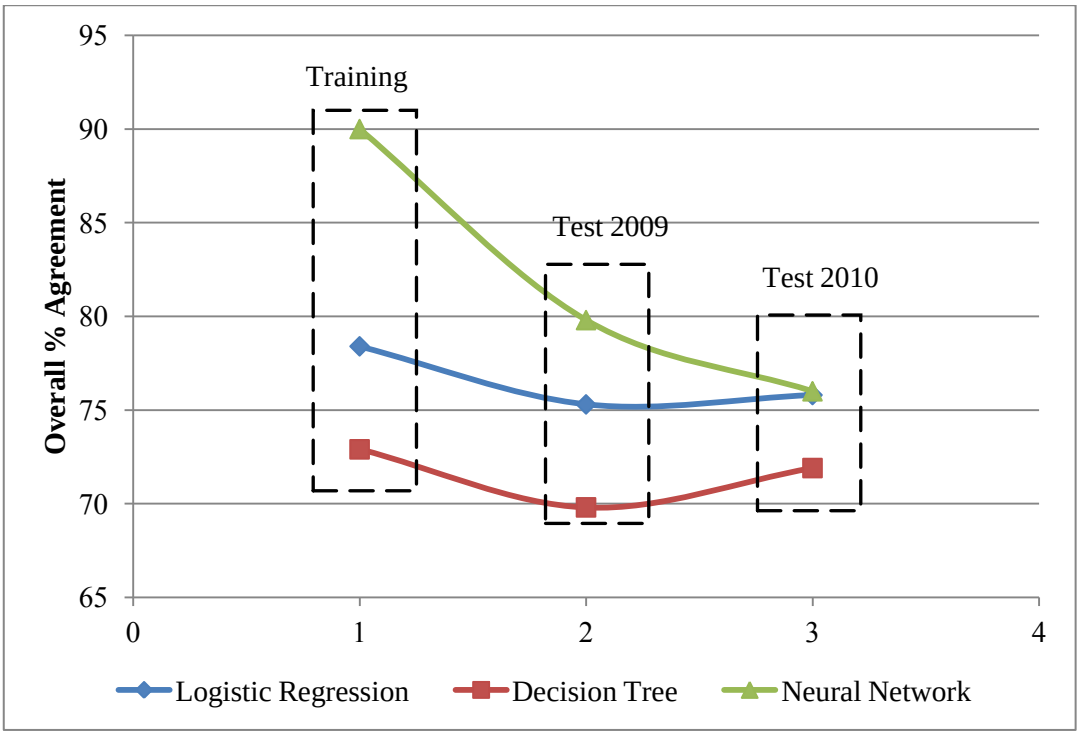

Figure 5.20: Overall prediction accuracy of prediction models in the revised analysis.

Table (5.20) presents the summary of prediction accuracy of the analysis when the models were applied at training sample size 800 and using the two types of sampling techniques, random and stratified sampling.

Table 5.20: Summary of the first analysis prediction accuracy results for sample size 800.

			% Agree		
			0	**1**	**Overall**
Logistic Regression	Random	Training	92.3	52.6	82.9
		Test 2009	91.3	50.6	81.8
		Test 2010	91.6	35.5	78.5
	Stratified	Training	93.3	54.9	84.0
		Test 2009	92.1	48.1	81.8
		Test 2010	94.9	27.5	79.2
Decision Tree	Random	Training	95.3	24.2	78.4
		Test 2009	94.5	23.6	77.9
		Test 2010	77.0	33.8	66.9
	Stratified	Training	96.4	29.5	80.3
		Test 2009	95.0	24.8	78.5
		Test 2010	95.8	16.3	77.3
Neural Network	Random	Training	96.6	85.8	94.0
		Test 2009	90.7	60.3	83.6
		Test 2010	90.4	37.8	78.1
	Stratified	Training	97.7	79.8	93.4
		Test 2009	91.3	55.0	82.8
		Test 2010	91.6	33.7	78.1

A comparison was conducted between the results obtained in Table (5.20) and the new results in Table (5.19) obtained using the new revised version of training data. The comparison resulted the following:

It is seen that the overall prediction accuracy of all models in the revised analysis were slightly lower than that of the first analysis. In the training data results in Table (5.20) of the first analysis, the maximum value of overall prediction accuracy was 94% and the minimum value was 78.4%. At the other side, the training data results in Table (5.19) of the revised training data analysis, the maximum value of overall prediction accuracy was 90% and the minimum value was 72.9%. The other results of overall prediction accuracy

for the testing data of both 2009 and 2010 in the two analyses were almost in the same range.

The prediction accuracy for the dependent variable values, “0” and “1”, using the prediction models were studied and compared and it is seen in Table (5.19) that the new revised training data succeeded to predict the two values of the dependent variable with comparable and high prediction ratios. In the first analysis in Table (5.20), if we take for example the prediction accuracy for the dependent variable values “0” and “1” when the decision tree model was applied on training data of stratified sample of size 800, the results were 95.3% and 24.2% respectively. On the other hand in the new revised training data analysis in Table (5.19) the prediction accuracy for the dependent variable values “0” and “1” when the decision tree model was applied on the training data were 70.2% and 75.6% respectively. This holds for the other remaining results in the two analysis in Table (5.19) and Table (5.20).

Fig. (5.21) and Fig. (5.22) plot a performance comparison of the overall prediction accuracy rate for the three models trained using different sized training data samples drawn using two sampling methods and tested using data of year 2009 and year 2010 respectively.

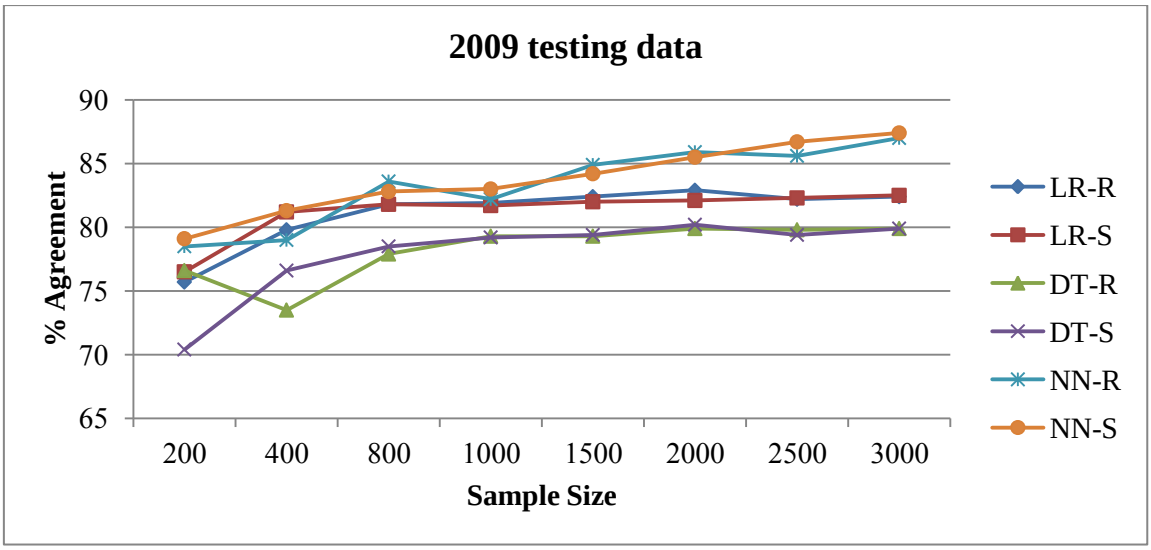

Figure 5.21: Overall prediction accuracy of prediction models within samples for 2009 testing data in the first analysis.

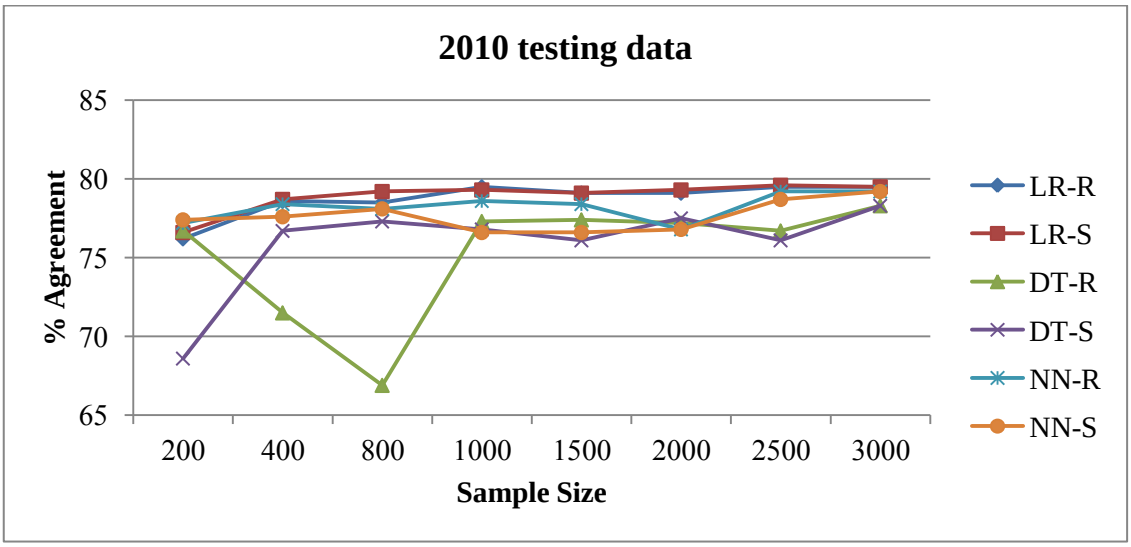

Figure 5.22: Overall prediction accuracy of prediction models within samples for 2010 testing data in the first analysis.

It is seen in Fig. (5.21) and Fig. (5.22) that the overall prediction accuracy for the three models was almost the same and the maximum difference between the lowest and highest nearly 8% and 12% in Fig. (5.21) and Fig. (5.22) respectively. If the results of decision tree were neglected at sample size 800 and smaller, then the overall prediction accuracy results for the three models were very close and the difference was very small. Also it is seen in Fig. (5.21) and Fig. (5.22) that when the training data sample size was 1000, the overall

prediction accuracy of the three models were too close and increasing the training data sample size more than 1000 didn’t significantly improve the prediction accuracy rate for any of the models.

It is seen from these results above that Neural Network algorithm outperformed the Decision Tree and Logistic Regression. This was due to the behavior of these algorithms in the classification process. The Logistic Regression depends on calculating the odds and probability of the desired output to be happened which result values between 0 and 1 forcing the algorithm to round the output values to be 0 or 1 only which consequently involves high error rate in classification process. So Logistic Regression cannot easily handle binary variables and it is not good for detecting interactions between variables. Due to this behavior Neural Network and Decision Tree are more powerful than Logistic Regression in modeling dependent variables with binary values also they can model categorical variables, with more than two discrete values, and they can handle variable interactions while Logistic Regression cannot do this [27], [28].

As seen in the background section Neural Networks is a simulation of highly interconnected neurons that provide models of data relationships that accept inputs, apply weighting coefficients and provide their output to be input to other neurons, forward or backward, that continue the process through the network to the final output and these steps are repeated in long and iterative process where the weights applied to each input at each neuron are adjusted to optimize the desired output. At the same time it is impossible to justify how decisions were made based on the output of the Neural Network and considered as a “black box”. On the contrary Decision Tree is easier than Neural Network; the resulting decisions can be explained easily and running faster than Neural Network, for training and classification, because Decision Tree, as a greedy algorithm, inherently throws away the inputs that it doesn't find useful, whereas a Neural Network will use them all.

Thus Decision Tree will find the solution faster than Neural Network and if it is lucky it will find an optimal solution which allows it to outperform the Neural Network in some datasets otherwise the Neural Network mostly outperform the Decision Tree [27], [30], [31], [32], [33].

Chapter Six

Conclusion

This research aimed to compare the classification performance of three statistical and data mining techniques, that are logistic regression, decision tree, and Neural Network, on different sized training data samples drawn from the PECS 2009 dataset using two sampling methods, simple random and stratified sampling methods, as training data to perform a performance comparison. To ensure that we had selected the suitable predictors, we performed a correlation analysis and selected the most correlated and significant variables that were related to the dependent variable (poverty status) (Appendix 2).

As a conclusion for this research, we can state that the sampling method has no effect on the prediction accuracy performance of any of the three prediction models. The sample size of the training data, which guides and controls the prediction process, does not have a vital role in increasing the prediction accuracy. On the other hand, for all of the prediction models in this study, the prediction accuracy performance maintained a steady state when the training data sample size reached 1000. This means that, in a huge datasets, to get a suitable prediction performance, no need to draw a big training dataset to train the prediction model and only 1000 records can do the training, which saves time, space and money. A tradeoff should be performed that whether the needed prediction accuracy is a high overall prediction accuracy rate; an adequate and comparable both “0” and “1”

dependent value prediction accuracy rate; or the needed is a high single “0” or “1” dependent values prediction accuracy rates. If high overall prediction accuracy is needed then the dependent variable values distribution in the training data should be skewed and the ratio of 0:1 occurrences (or 1:0) should be at least 2:1 or larger. This hold also if the requested high prediction accuracy is one of the two dependent variable values, not both, then it’s distribution in the training data should be at least two occurrences or more against to one occurrence for the other value. An example of this case is the breast cancer diagnosing in women that high prediction accuracy is needed to check if the patient is infected. If both dependent variable values are requested to be predicted in comparable prediction accuracy rate, then the training data should not be skewed and the ratio of dependent variable values occurrences should be equal and no more than 1:1. This holds for all of the three prediction models.

A general conclusion could be stated that Neural Network outperformed the other two models. These conclusions contradict with the results and conclusion of Lahiri R. (2006) [1] about the Neural Network failure to predict the individual dependent variable’s values. In this study it is seen in the *Revised Training Data Results* section that the Neural Network succeeded and outperformed the logistic regression and decision tree models in predicting the individual values of the dependent variable values, “0” and “1”, (Table 5.19), (Fig. 5.19, Fig. 5.20, Fig. 5.21 and Fig. 5.22). We expected that the result of this difference was the data content and quality. It is worth mentioning that because Logistic Regression algorithm produces a high level of error rate because of the probability output values rounding that put this algorithm away from the comparison between Decision Tree an Neural Network which delivers more accurate prediction values. Because the greedy Decision Tree algorithm depends on batch-learning and inherently throws away the inputs

that it doesn't find useful, whereas a Neural Network use them all, Neural Network outperformed the Decision Tree.

As a future work, we recommend performing the same study to compare the prediction performance of the logistic regression, decision tree and Neural Network within different sized samples and different dependent variable values distributions. Another future work we recommend to study and survey a collection of classifiers over a set of diversity of datasets and produce a general framework of prediction models by mapping each classifier to the suitable dataset type where the classifier outperforms the other classifiers.

References

[1] Lahiri R. (December 2006): Comparison of Data Mining and Statistical Techniques for Classification Model – A Thesis submitted to the graduate faculty of the Louisiana State University in partial fulfilment of the requirements for the degree of Master of Science in The Department of Information Systems & Decision Sciences.

[2] Delen D., Walker G., and Kadam A. (2005 Jun): Predicting Breast Cancer Survivability: A Comparison of Three Data Mining Methods – Artificial Intelligence in Medicine – 34(2):113–27.

[3] Bellaachia A. and Guven E. (2006): Predicting Breast Cancer Survivability Using Data Mining Techniques – Ninth Workshop on Mining Scientific and Engineering Datasets in conjunction with the Sixth SIAM International Conference on Data Mining.

[4] M. Panda and M. R. Patra (2008): A Comparative Study Of Data Mining Algorithms For Network Intrusion Detection – proc. of ICETET, India.pp.504–507. IEEE Xplore.

[5] Amooee G., Minaei–Bidgoli B. and Bagheri–Dehnavi M. (November 2011): A Comparison Between Data Mining Prediction Algorithms for Fault Detection (Case study: Ahanpishegan co.) – IJCSI International Journal of Computer Science Issues – Vol. 8, Issue 6, No 3.

[6] Ho Yu C., DiGangi S., Jannasch–Pennell A., and Kaprolet C. (2010): A Data Mining Approach for Identifying Predictors of Student Retention from Sophomore to Junior Year – Journal of Data Science 8, 307–325.

[7] Kumari M. and Godara S. (June 2011): Comparative Study of Data Mining Classification Methods in Cardiovascular Disease Prediction – International Journal of Computer Science and Technology (IJCST) Vol. 2, Issue 2.

[8] Shailesh K R et. al. (2011): Comparison of Different Data Mining Techniques to Predict Hospital Length of Stay, JPBMS, 7 (15).

[9] Ibrahim Z. and Rusli D. (September 2007): Predicting Students' Academic Performance, Comparing Artificial Neural Network, Decision Tree and Linear Regression – 21st Annual SAS Malaysia Forum, 5th.

[10] Nancy P. and Geetha Ramani R. (October 2011): A Comparison on Performance of Data Mining Algorithms in Classification of Social Network Data – International Journal of Computer Applications (0975–8887) Volume 32– No.8.

[11] C. Deepa, K. Sathiya Kumari, and V. Pream Sudha (2011): A Tree Based Model for High Performance Concrete Mix Design – International Journal of Engineering Science and Technology Vol. 2(9), 4640–4646.

[12] S. Shanthi and R. Geetha Ramani (December 2011): Classification of Vehicle Collision Patterns in Road Accidents using Data Mining Algorithms – International Journal of Computer Applications (0975–8887) Volume 35–No.12.

[13] Palestinian Central Bureau of Statistics (2010) – Levels of living in the Palestinian Territory. Final Report (January 2009– January 2010) – Ramallah – Palestine.

[14] Wikipedia, the free encyclopedia (2012): (http://en.wikipedia.org/wiki/Logistic_regression).

[15] Han J. and Kamber M. (2007): Data Mining: Concepts and Techniques – Second Edition, Elsevier Inc.

[16] DTREG (2012): Multilayer Perceptron Neural Networks, A Brief History of Neural Networks. (http://www.dtreg.com/mlfn.htm).

[17] Smith L. (April 2003): An Introduction to Neural Networks. (http://www.cs.stir.ac.uk/~lss/NNIntro/InvSlides.html#what).
[18] NeruoSolutions (2012): Introduction to Neural Networks and NeuroSolutions – Video Presentation. (http://www.nddownloads1.com/videos/nns_and_neurosolutions/nns_and_neurosolutions.html)
[19] SPSS Inc. (2012): PASW Neural Networks 18 – statistical package for the social sciences. Chicago, USA. (www.spss.com).
[20] Mitchell T. (1997): Machine Learning, McGraw Hill.
[21] Hajek M. (2005): Neural Networks.
[22] NeruoSolutions (2012): What is Neural Networks. (http://www.nd.com/welcome/whatisnn.htm)
[23] Harb A. and Jayousi R. (2012): A Comparative Study of Statistical and Data Mining Algorithms for Prediction Performance-proceedings of the International Conference on Information & Communications Technology (ICICT'2012).
[24] Harb A. and Jayousi R. (2012): Comparing Neural Network Algorithm Performance Using SPSS and Neurosolutions -proceedings of the 13th International Arab Conference on Information Technology (ACIT'2012).
[25] Harb A. and Jayousi R. (2013): Impact of Data Size on Data Mining Prediction Accuracy-the 6th International Conference on Information Technology (ICIT'2013).
[26] Palestinian Central Bureau of Statistics (PCBS). (http://www.pcbs.gov.ps)
[27] Jae H. Song, Santosh S. Venkatesh, Emily A. Conant, Peter H. Arger, Chandra M. Sehgal (2005): Comparative Analysis of Logistic Regression and Artificial Neural Network for Computer-Aided Diagnosis of Breast Masses- Academic Radiology, Vol 12, No 4, April 2005
[28] DTREG (2012): Decision Trees Compared to Regression and Neural Networks. (http://www.dtreg.com/othermethods.htm).
[29] Barnaghi P. M., Sahzabi V. A. and Abu Bakar A.(2012): A Comparative Study for Various Methods of Classification- International Conference on Information and Computer Networks (ICICN 2012)
[30] Khemphila A., Boonjing V. (2010): Comparing performances of logistic regression, decision trees, and Neural Networks for classifying heart disease patients- Computer Information Systems and Industrial Management Applications (CISIM), 2010 International Conference on , vol., no., pp.193-198, 8-10 Oct. 2010
[31] Bouzida Y. and Cuppens F. (2006): Neural networks vs. decision trees for intrusion detection-IEEE / IST Workshop on Monitoring, Attack Detection and Mitigation 2006.
[32] Chandra R., Chaudhary K. and Kumar A. (2007): The Combination and Comparison of Neural Networks with Decision Trees for Wine Classification- School of sciences and technology, University of Fiji, 2007
[33] Kumar K. (2012): Knowledge Extraction From Trained Neural Networks- International Journal of Information & Network Security (IJINS), Vol.1, No.4, October 2012, pp. 282~293

Appendix 1: Data Dictionary of original PECS' data (2009 and 2010).

File Name: Identification *All variables are numeric*

Variable	Label	Values	Measure
ID00	Questionnaire serial no. in sample		Scale
Area	Gaza Strip and West Bank Areas	1: North of West Bank 2: Middle of West Bank 3: South of West Bank 4: Gaza Strip	Scale
Loc_type	Location Type	1: urban 2: rural 3: camp	Scale
IR04_male	Number of males in household		Scale
IR04_female	Number of females in household		Scale
Region	Area	1: west bank , 2: Gaza	Scale
RW	Relative weight		Scale

File Name: Monthly Income *All variables are numeric*

Variable	Label	Values	Measure
ID00	Questionnaire serial no. in sample		Scale
RW	Relative Weight		Scale
Income	Total Household Monthly Income in Israeli Shekel		Scale

File Name: Main Groups *All variables are numeric*

Variable	Label	Values	Measure
ID00	Questionnaire serial no. in sample		Scale
Amount of consumption and expenditure on groups…. Grp1: Bread and Cereals : Grp30: Social protection			
cons	total consumption		Scale
exp	total expenditure		Scale
rw	relative weight		Scale

File Name: Roster *All variables are numeric*

Variable	Label	Values	Measure
ID00	Questionnaire serial no. in sample		Scale
D1	Line no. of member		Scale
D3	Relationship of member to the head of household	1: Head of HH 2: Husband/ wife 3: Son/daughter 4: Father/mother 5: Brother/sister 6: Grandfather/mother 7: Grandchild 8: Daughter/son in law 9: Other relatives 10: Other	Nominal

File Name: Roster *All variables are numeric*

Variable	Label	Values	Measure
D4	Sex	1: Male , 2: Female	Nominal
D5	Age		Scale
D6	Refugee Status	1: Refugee , 2: Not Refugee	Nominal
D11	Does he has medical insurance?	1: Yes , 2: No	Nominal
D14	Education Attendance	1: Currently attending school 2: Attended school at any time and left before completing level 3: Attended school and graduated 4: Never attended school	Nominal
D15	Number of education years that		Scale
D16	Educational Status	1: Illiterate 2: Can read and write 3: Elementary 4: Preparatory 5: Secondary 6: Associate diploma 7: Bachelor 8: High diploma 9: Master 10: Ph.D	Nominal

File Name: Roster *All variables are numeric*

Variable	Label	Values	Measure
D17	What is the main reason for dropping out of school (for persons 5 years and more)?	1: Unwillingness for academic education 2: Unwillingness for co-education 3: Frequent repetition 4: Not interested in study 5: Bad economic situation of the family 6: Existing family problems 7: Caring for members of the family 8: Marriage 9: Sickness 10: Disability 11: No school nearby 12: Mistreatment at school 13: Security situation 14: Dismissal from school because of exceeding the legal age 15: Other	Nominal

File Name: Roster *All variables are numeric*

Variable	Label	Values	Measure
D18	Work Status during the past week (for persons aged 7 years and over)	1: Employed from 1-14 hours 2: Employed 15-34 hours 3: 35 hours and over 4: Looked for work last week 5: Did not looked for work because of frustration 6: looked for work last week 7: Did not looked for work because of frustration 8: Full time student 9: Housewife 10: Unable to work 11: has revenue 12: other	Nominal
D19	Work Status for persons aged 7 years and over	1: Employer 2: Self employed 3: Unpaid Employee 4: work for regular wage 5: work for irregular wage	Nominal

File Name: Roster *All variables are numeric*

Variable	Label	Values	Measure
D20	Place of Work	1: In the same Locality 2: In the same governorate 3: In other Governorate 4: Israel/ Settlements 5: Abroad	Nominal
D21	Main Occupation Describe main tasks for coding	1: Legislators, Senior Officials & Managers 2: Professionals, Technicians, Associates and Clerks 3: Service, Shop & Market Workers 4: Skilled Agricultural & Fishery Workers 5: Craft and Related Trade Workers 6: Plant & Machine Operators & Assemblers 7: Elementary Occupations	Scale
D22	Economic Activity	1: Agriculture, fishing and forestry 2: Mining, quarrying and manufacturing	Scale

File Name: Roster *All variables are numeric*

Variable	Label	Values	Measure
		3: Construction 4: Commerce, restaurants and hotels 5: Transportation, storage and communication 6: Services and other branches	
D23	Sector	1: National private inside establishments 2: National private outside establishments 3: Foreign private inside establishments 4: Foreign private outside establishments 5: National government 6: Foreign government 7: Charitable association 8: UNRWA 9: International organization	Nominal
D24	Does person have another work	1: Yes 2: No	Nominal
D25	Number of working months during the		Nominal

File Name: Roster *All variables are numeric*

Variable	Label	Values	Measure
	year. If not working during the year, write 00		
D26	Marital Status(for persons 12 years and over)	1: Never married 2: Legally married 3: Currently married 4: Divorced 5: Widowed 6: Separated	Nominal
RW	Relative weight		Scale

File Name: Dwelling Characteristics *All variables are numeric*

Variable	Label	Values	Measure
ID00	Questionnaire serial no. in sample	None	Scale
H1	Type of housing unit	1: Villa 2: House 3: Apartment 4: Separate Room 5: Tent 6: Marginal 7: Others	Scale

File Name: Dwelling Characteristics *All variables are numeric*

Variable	Label	Values	Measure
H2	Tenure of the housing unit	1: Owned 2: Rented no furniture 3: Rented with furniture 4: Without payment 5: For work 6: Others (specify)	Scale
H3	What is the main material used in building outside walls of housing unit		Scale
H4	usage of housing unit	1: for residence 2: residence & work	Scale
H5	Number rooms are there in dwelling		Scale
H6	No. of sleeping rooms are used in dwelling		Scale
H7A	The monthly rent		Scale
H7B	specify type of currency	1: shekel 2: Jordanian Dinar 3: Dollar	Scale
H8A	The estimated monthly rent		Scale
H8B	Specify type of currency	1: shekel 2: Jordanian Dinar 3: Dollar	Scale

File Name: Dwelling Characteristics *All variables are numeric*

Variable	Label	Values	Measure
H9A	Connection to Public Networks -water	1: Local Public network 2: Israeli network 3: rain water 4: Bridges 5: Tank 6: other	Scale
H9B	Connection to Public Networks - electricity	1: Public network 2: Private generator 3: No electricity	Scale
H9C	Connection to Public Networks -sewage	1: Public Sewage System 2: hole absorption 3: Cesspit 4: No Sewage System	Scale
H10	Availability of a kitchen	1: Kitchen with Piped Water 2: Kitchen without Piped Water 3: No Kitchen	Scale
H11	Availability of a bathroom	1: Bathroom with Piped Water 2: Bathroom without Piped Water 3: No Bathroom	Scale

File Name: Dwelling Characteristics *All variables are numeric*

Variable	Label	Values	Measure
H12	Availability of a toilet (wc)	1: Toilet with Piped Water 2: Toilet without Piped Water 3: No Toilet	Scale
H13_1	Main source of energy -cooking	1: Gas 2: Kerosene 3: Electricity 4: Wood 5: Other / specify	Scale
H13_2	Main source of energy -heating	0: No exist 1: gas 2: Kerosene 3: Electricity 4: Wood/coal 5: Other/ specify	Scale
H13_3	Main source of energy - Conditioner	0: No exist 1: Electricity 2: Other/ specify	Scale

File Name: Dwelling Characteristics *All variables are numeric*

Variable	Label	Values	Measure
H13_4	Main source of energy - Oven	0: No exist 1: gas 2: Electricity 3: Wood 4: olive cake 5: coal 6: Other/ specify	Scale
H13_5	Main source of energy - Water heater	1: Sun 2: Gas 3: Kerosene 4: Electricity 5: Wood 6: Coal 7: solar 8: Other/ specify	Scale
H14_1	Dampness	1: Yes , 2: No	Scale
H14_2	Cold and	1: Yes , 2: No	Scale
H14_5	difficult heating in winter	1: Yes , 2: No	Scale
H14_3	Poor ventilation	1: Yes , 2: No	Scale
H14_4	High heat in summer	1: Yes , 2: No	Scale
H18_1	Smoke, exhaust from cars	1: Yes , 2: No	Scale
H18_2	Smoke, exhaust from industry	1: Yes , 2: No	Scale
H18_3	Odors resulting from animals	1: Yes , 2: No	Scale

File Name: Dwelling Characteristics *All variables are numeric*

Variable	Label	Values	Measure
H18_4	Odors resulting from sewage system water	1: Yes , 2: No	Scale
H18_5	Odors resulting from garbage	1: Yes , 2: No	Scale
H18_6	General dust	1: Yes , 2: No	Scale
H18_7	Dust or smells resulting from other sources	1: Yes , 2: No	Scale
H18_8	Noise	1: Yes , 2: No	Scale
H19	the method for removing garbage	1: Collected by sanitation worker 2: Thrown in nearby garbage container 3: Thrown randomly 4: Thrown in garbage area 5: Burned 6: Used for specific things 7: Other / specify	Scale
H20_1	Public transportation	1: Less than 1 km 2: 1-5 km 3: More than 5 km	Scale
H20_2	Private doctor clinic	1: Less than 1 km 2: 1-5 km 3: More than 5 km	Scale

File Name: Dwelling Characteristics *All variables are numeric*

Variable	Label	Values	Measure
H20_3	Health center	1: Less than 1 km 2: 1-5 km 3: More than 5 km	Scale
H20_4	Hospital	1: Less than 1 km 2: 1-5 km 3: More than 5 km	Scale
H20_5	Elementary/ Secondary school	1: Less than 1 km 2: 1-5 km 3: More than 5 km	Scale
H20_6	Mother and child health central	1: Less than 1 km 2: 1-5 km 3: More than 5 km	Scale
H21_1	Private Car	1: Yes , 2: No	Scale
H21_2	Refrigerator	1: Yes , 2: No	Scale
H21_3	Solar Boiler	1: Yes , 2: No	Scale
H21_4	Washing Machine	1: Yes , 2: No	Scale
H21_5	Cooking stove	1: Yes , 2: No	Scale
H21_6	Dish washer	1: Yes , 2: No	Scale
H21_7	Central heating	1: Yes , 2: No	Scale
H21_8	Vacuum cleaner	1: Yes , 2: No	Scale
H21_9	Dehumidifier	1: Yes , 2: No	Scale
H21_10	Home library	1: Yes , 2: No	Scale
H21_11	T.V	1: Yes , 2: No	Scale

File Name: Dwelling Characteristics *All variables are numeric*

Variable	Label	Values	Measure
H21_12	Video/DVD	1: Yes , 2: No	Scale
H21_13	Phone line	1: Yes , 2: No	Scale
H21_14	Jawwal	1: Yes , 2: No	Scale
H21_15	Mobile Israel	1: Yes , 2: No	Scale
H21_16	Computer	1: Yes , 2: No	Scale
H21_17	Satellite	1: Yes , 2: No	Scale
H21_18	Microwave	1: Yes , 2: No	Scale
H21_19	Radio/Recoeder	1: Yes , 2: No	Scale
H21_20	Filter	1: Yes , 2: No	Scale
H21_21	Other	1: Yes , 2: No	Scale
H22_1	Animals for transportation	1: Yes , 2: No	Scale
H22_2	Taxi	1: Yes , 2: No	Scale
H22_7	Truck	1: Yes , 2: No	Scale
H22_3	Tractor	1: Yes , 2: No	Scale
H22_4	Container water	1: Yes , 2: No	Scale
H22_8	Tailoring machine	1: Yes , 2: No	Scale
H22_5	Craft jobs	1: Yes , 2: No	Scale
H22_6	Trade jobs	1: Yes , 2: No	Scale
H22_9	Other	1: Yes , 2: No	Scale
H23	Household Main Source of Income	1: Household Business 2: Wages & Salaries 3: Remittances in Cash\ Other Sources	Scale

File Name: Dwelling Characteristics *All variables are numeric*

Variable	Label	Values	Measure
H24	family have a agricural	1: Yes , 2: No	Scale
H32	the household have animal holdings (Cattle, Sheep and Goats, Poultry, Horses and Mules, Beehives)	1: Yes , 2: No	Scale
RW	Relative Weight		Scale

Appendix 2: Data Dictionary of the final data of the research.

File Name: Household Characteristics *All variables are numeric*

Name	Label	Values	Measure
PovStat	Poverty status	0: Not poor , 1: Poor	Nominal
HHDens	Household density = Hsize / HRooms		Scale
HSize	Household size		Scale
HRooms	Number rooms are there in dwelling		Scale
UnEmployed	Number of unemployed persons in household		Scale
HMale	Number of males in household		Scale
Children1	Number of children less than 6		Scale
Children2	Number of children between 6 and 11		Scale
Children3	Number of children between 12 and 15		Scale
Stone	The main material used in building outside walls of housing unit is stone	0: not Stone 1: Stone or old stone	Nominal
AreaGza	Gaza strip	0: others 1: Gaza Strip	Nominal
AreaMid	Middle West Bank	0: others 1: Middle West Bank	Nominal
Household's head information			
WorkMnths	Number of working months during the year		Scale
WorkType	Work type	0: = Employer or Work for	Nominal

File Name: Household Characteristics *All variables are numeric*

Name	Label	Values	Measure
		regular wage 1: else	
Occup	Main Occupation	0: others 1: Elementary Occupations	Nominal
EduStatus	Education status is diploma and above	0: other 1: diploma and above	Nominal
Availability of services in the household			
Micowv	Microwave	0: YES , 1: NO	Nominal
VacCln	Vacuum cleaner	0: YES , 1: NO	Nominal
Compu	Computer	0: YES , 1: NO	Nominal
Phone	Phone line	0: YES , 1: NO	Nominal
VidDVD	Video/DVD	0: YES , 1: NO	Nominal
Car	Private car	0: YES , 1: NO	Nominal
CokStov	Cooking stove	0: YES , 1: NO	Nominal
Radio	Radio/Recorder	0: YES , 1: NO	Nominal
Refrig	Refrigerator	0: YES , 1: NO	Nominal
HotWtrSrc	Main source of energy for water heater	0: Sun or Gas or Electricity 1: others	Nominal
HLib	Home library	0: YES , 1: NO	Nominal
IsrMob	Israeli mobile	0: YES , 1: NO	Nominal
WshMach	Washing machine	0: YES , 1: NO	Nominal
SBoiler	Solar boiler	0: YES , 1: NO	Nominal
CondSrc	Main source of energy for air	0: Electricity	Nominal

File Name: Household Characteristics *All variables are numeric*

Name	Label	Values	Measure
	conditioner	1: No exist or others	
OvnSrc	Main source of energy for oven	0: Gas or Electricity or Not exist 1: others	Nominal
Satlt	Satellite	0: YES , 1: NO	Nominal
Kchn	Availability of a kitchen	0: Kitchen with piped water 1: Kitchen without piped water or no kitchen	Nominal
TV	T.V	0: YES , 1: NO	Nominal
Dehum	Dehumidifier	0: YES , 1: NO	Nominal
WtrFltr	Water filter	0: YES , 1: NO	Nominal
BthRom	Availability of a bathroom	0: Bathroom with piped water 1: Bathroom without piped water or no bathroom	Nominal
CenHet	Central heating	0: YES , 1: NO	Nominal
OtherEquip	Other equipment	0: YES , 1: NO	Nominal